# Design and analysis of fixed and adaptive sigma-delta modulators.

Yu, Jie

# Design and Analysis of Fixed and Adaptive Sigma-Delta Modulators

by

*Jie Yu*

A thesis submitted for the degree of

Doctor of Philosophy

Department of Electronic and Electrical Engineering

King's College, University of London

September 1992

# ABSTRACT

Oversampled sigma-delta modulation has become an increasingly popular means of A/D and D/A conversion in recent years, particularly in digital audio systems. Low bit quantisation is used in conjunction with a carefully designed loop filter and oversampling to achieve high resolution performance. This thesis reports on efforts to design and analyse sigma-delta modulators and to investigate the performance of adaptive sigma-delta modulation systems.

Conventional design methods for a one-bit sigma-delta modulator assume an additive white noise model and therefore are based on linear system theory. It is found that in some cases the expected signal-to-noise ratio and stability properties could not be achieved. This is because of the severe nonlinearity of the one-bit quantiser inside the system loop. An optimisation method which does not depend on the additive white noise model is proposed in order to determine the coefficients of the loop filter which maximise signal-to-noise ratio. Also, the optimal quantisation level is determined by analysis and computer simulation.

Once designed, a conventional sigma-delta modulator operates with fixed coefficients and fixed quantisation step size. Very little work has been carried out on adaptive sigma-delta modulators. In this thesis, the idea of adaptive quantisation is applied to increase the dynamic range of the modulators. Based on the concept of an equivalent quantiser, block backward adaptation logic is employed. It is shown that the dynamic range can be increased dramatically. A fast attack time, but a slow release time, are established so as to tailor the system to music signals. The oversampling ratio or the order of the loop filter can be reduced while maintaining the same signal-to-noise ratio for small signals at the expense of an increase in noise for the rarely occurring large signals. Also, the idea of adaptive loop filter is proposed. The use of the adaptive filter is shown to lead to a slight increase in the dynamic range of the system.

# ACKNOWLEDGEMENTS

I would like to express my appreciation to Mr R.E. Hawken for his assistance and supervision in this project. A special word of gratitude to Dr M.B. Sandler for suggesting the topic of study and his many useful discussions.

My early education in Beijing University of Aeronautics and Astronautics, and in Tsinghua University, Beijing, China, has given me the fundamental knowledge and the basic skills, which are very helpful for this research work. Among my previous University teachers, I would like to thank Professors Yingjie Yu, Ziming Ding, Chongxi Feng, Yasheng Qian, and Qinglin Zhu. They played important roles as mentors as well as my friends and gave me great encouragement during my previous education.

My thanks also go to many of my colleagues for their help and discussions during the years of my study. Among them are Hamid Dabirikhah, who often gave great encouragement and many helps, Jason Goldberg, who gave many useful discussions and great help in English language, Allan Paul, who provided a very good environment in Sunwork Station, which made the simulations possible for this project. Other people with whom I have had good fortune to collaborate and learn from include Rod Hiorns, Rob Bowman, Mike Waters, Jonathan Machenzie, Fei Xia and many others.

I gratefully acknowledge the financial support of the KC Wong Scholarship from King's College and by the Chinese Educational Committee.

*To My Parents*

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND GLOSSARY TERMS

N - Oversampling ratio

L - Number of octaves of oversampling ratio; $N=2^L$

B - Number of bits of a quantiser

n - Number of stages of a comb filter or order of the loop filter in a sigma-delta modulator

d - Quantisation level

k - Time-domain index

$f_b$ - Nyquist sampling frequency, i.e., twice of the signal band

$f_s$ - Sampling frequency: $f_s = Nf_b$

G(z) - Loop filter in the forward path of a sigma-delta modulator

Q(u) - Function of a quantiser with the input u

$Q_x(x)$ - Equivalent quantiser with the input x

*Quantisation level* - The absolute value of the output of one-bit quantiser

*Decimator* - In this thesis, a decimator is defined as a system which only reduces the sampling rate of a digital signal; it does not include low-pass filtering.

*Equivalent quantiser* - An equivalent quantiser is a single quantiser used to model the complete sigma-delta modulation system, that is, the input to the equivalent quantiser is the input to the sigma-delta modulator while its output is the output of the demodulator.

# 1

# INTRODUCTION

Although the principle for binary-coded pulse code modulation (PCM) was established around 1937 [1], digital conversion and storage technology were not sufficiently advanced to challenge analogue techniques until 1960. Several technical advances in the 1960's began to facilitate research in the new area of digital signal processing. The most relevant of these developments were the emergence of low-cost, high-speed digital circuits, mini-computers, and digital instrumentation technology. These great achievements of digital signal processing technology have revolutionised signal processing in the areas of communication, speech processing, image processing, digital audio, and control. The benefits of digital representation are many and well known [2]. Perhaps most significant is the fact that digital signals are less sensitive than analogue signals to noise. They are easy to process, regenerate, and store. Today, an analogue signal is typically processed in digital format by using discrete time sampling and discrete magnitude quantising, which is called analogue-to-digital conversion (ADC). The conversion of the processed digital signal back to the analogue waveform by means of digital-to-analogue converters (DACs) takes place only after the digital signal processing has been completed. Common examples are digital telephone systems and

modern digital audio systems like compact disk (CD) players and digital audio tape systems. Thus, the analogue-to-digital and digital-to-analogue interfaces have become crucial links between the analogue signal and its digital processor and, as a result, it has been a topic of extensive study for more than 40 years.

## 1.1 A/D and D/A conversion

Analogue-to-digital and digital-to-analogue conversions play key roles towards "all-digital" realisation. The generation of a digital signal from an analogue one is a simple process conceptually. The block diagram of a basic A/D converter is shown in Fig. 1-1. First, the frequency band of the signal must be limited by an anti-aliasing, i.e., low-pass filter. The sample-and-hold circuit will take samples periodically from the analogue input signal with a sampling frequency greater than twice the highest frequency given by the bandwidth of the signal according to Nyquist's sampling theorem and maintain the instantaneously obtained amplitude constant for a certain period. During this hold period the A/D converter performs its main functions, namely quantising and encoding.

To create an analogue signal from a digital sequence, these steps are reversed as shown in Fig. 1-2. Upon applying the digital signal with limited wordlength at the input of a D/A converter, a single discrete analogue value is obtained at its output. This single value replaces an infinite number of originally continuous values of the original signal, i.e., it represents the original signal plus some quantising noise. The next component in typical DAC systems is a special sample-and-hold amplifier which prevents the internal switching glitches of a DAC from appearing in the analogue

output signal. It does this by holding the previous DAC voltage at its output for a certain period of time, while the new DAC voltage settles at its input. The role of the reconstruction filter is to remove the energy in the spectral images of the baseband which, due to the sampling process, exist at multiples of sampling frequency: $f_S=1/T$.

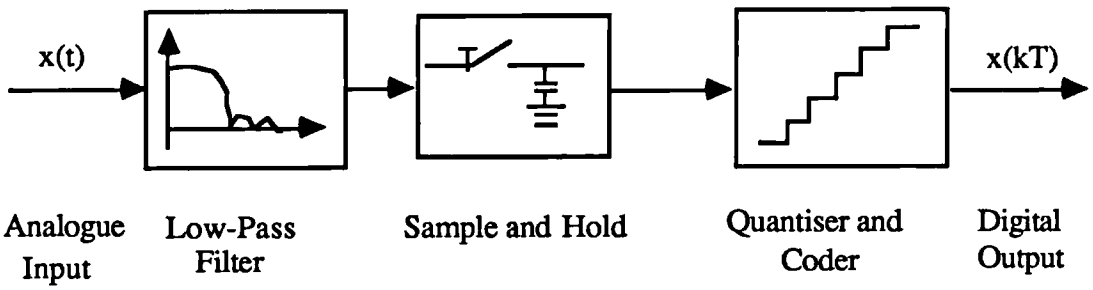



Fig. 1-1 Main components in an analogue-to-digital conversion system

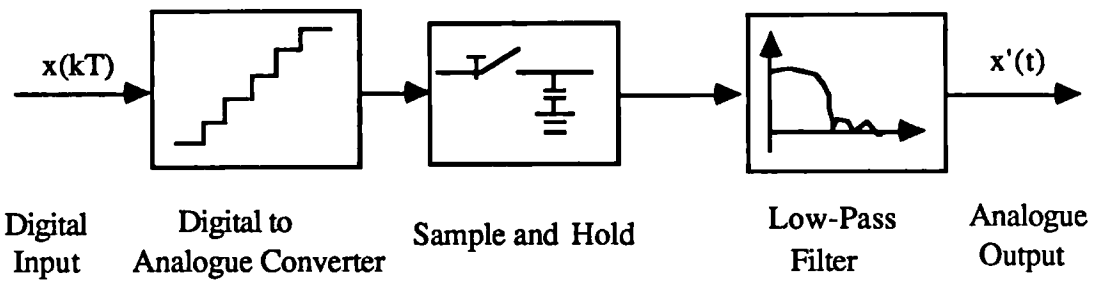



Fig. 1-2 A digital-to-analogue conversion system

## 1.2 Major techniques for A/D implementation

There are three major conventional techniques for implementing analogue to digital conversion: successive-approximation, ramp-comparison, and flash ADCs. Ramp-comparison converters are simple, but involve a large number of sequential operations that restrict their use to relatively slow speed applications. Flash converters are complex but fast, and therefore suitable for high speed applications. Successive approximation converters provide a compromise between speed and complexity.

Because of the increasing complexity of modern systems, it is desirable to have the analogue and digital interfaces integrated with other system modules so that the entire system can be fabricated on the same semiconductor chip. Consider the case of flash converters. An R-bit flash ADC requires $2^R$-1 threshold elements, each representing a decision level. Each sample of the input signal is compared with the thresholds and then an output symbol is decided based on the result of the comparisons. Although this method is in principle simple and straightforward, a problem arises when the quantiser circuit is built. Suppose we want to build a 16-bit quantiser on a VLSI chip with a dynamic range of 0-5 Volts. Then, we need to build 65535 comparators onto the chip with threshold levels of neighbouring comparators which differ by only 76 $\mu$V. Therefore, the precisions required by modern digital audio systems, say, 16- to 24-bit resolution, make the VLSI implementation very difficult. Furthermore, the conventional converters require a sharp analogue anti-aliasing or anti-imaging filter, and an analogue sample and hold circuit.

Oversampled A/D converters achieving high resolution by using a low resolution quantiser have recently received considerable attention. Fig. 1-3 demonstrates the basic

idea. Sigma-delta modulation is often used for this type of converter. The input signal is sampled at a rate many times faster than the required Nyquist rate. The sampled input signal is then quantised by a low resolution B-bit quantiser inside a feedback loop containing a low-pass analogue filter. The analogue filter combining with the feedback loop shapes the large quantisation noise produced by the coarse quantiser, moving most of its energy to frequencies above the desired signal band, or, baseband. A digital low-pass filter then removes the out-of-band shaped noise and decimates the high rate low-bit (B-bit) output stream so as to produce the final high resolution M bit digital signal at the Nyquist frequency, where M is much larger than B. The same idea has been used for interpolated D/A converters. The resulting converters release the need for precise analogue anti-aliasing and anti-imaging filters. They are more robust against circuit imperfections because there are fewer bits used in the quantiser and the bulk of the signal processing is performed in digital circuitry. Moreover, because they sample the analogue input signal at well above the Nyquist rate, precision sample-and-hold circuitry is unnecessary. Therefore, they are well suited to VLSI implementation.

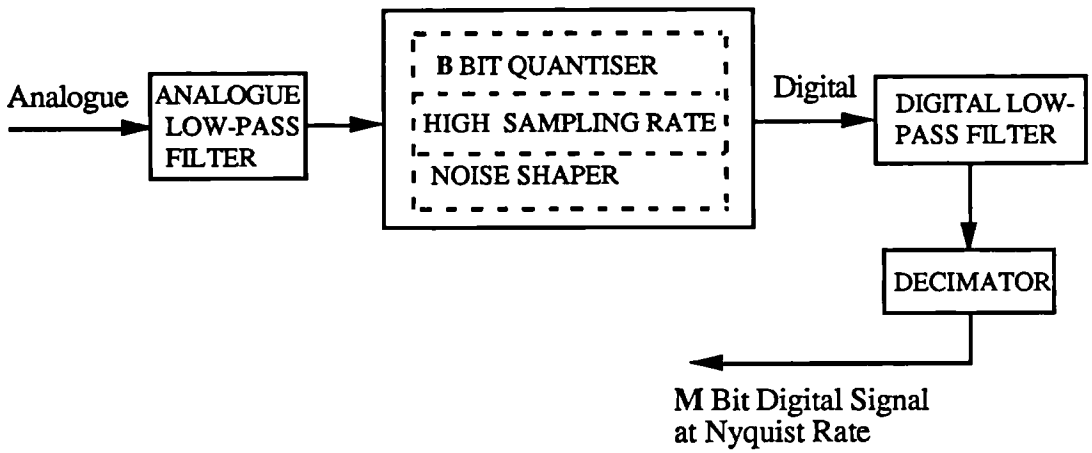



Fig. 1-3 Block diagram which demonstrates the basic idea of an oversampled A/D converter

## 1.3 Brief history and motivation

Sigma-delta modulation (SDM) was first formally proposed by Inose Yasuda and Murakami in around 1962 for communication and telemetering systems [3][4]. The initial idea was to pre-emphasise the low frequencies in the input by simply integrating the input prior to delta modulation (DM) coding in order to transmit dc and very low frequency components. The procedure also simplifies DM decoding because conventional integration in DM decoding can be removed and only a low-pass filter is needed. Compared with DM, the overload characteristic of the SDM is frequency-independent. This property makes SDM an ideal device for A/D conversion [5].

The modern popularity of sigma-delta modulation is mainly attributed to the work of Candy and his colleagues. Candy studied the limit cycle structure of the sigma-delta modulator in 1974. It was pointed out [6] that a SDM uses the limit cycles in a coarsely quantising feedback coder to give a precise determination of the average input value. In 1981, Candy and Benjamin [7] carried out an analysis based on a simple approximate continuous-time model to obtain an understanding of the structure of quantisation noise from sigma-delta modulation. They stated that the modulation noise is highly correlated with the amplitude of a dc input. The measurement of baseband rms noise over the range of dc inputs showed that noise is largest when the modulator is biased near the ends of its range and next largest near the centre. Peaks of noise occur in pairs and are most prominent when the sampling rate is large.

A sigma-delta modulator is a nonlinear feedback system. A long-standing problem with such nonlinear systems has been the difficulty in analysing their exact behaviour. There are basically two different approaches for analysing SDMs: the

additive white noise source approximation, and the rigorous or direct techniques. In the first approach which was the most common method, one tries to approximate the quantiser noise by choosing an input-independent additive noise source having a similar long-term average behaviour to the actual quantiser noise, e.g., the same long-term sample distribution and power spectrum. The simplest noise model is white noise with a uniform distribution. Under such an approximation the nonlinear SDM is modelled as a linear system, and the performance can easily be derived by using well-known linear system techniques such as standard Fourier analysis. Some of the properties derived by using this approach agree reasonably well with simulation results [8][9].

However, this model is not accurate enough for lower bit quantisers, especially for single bit quantisers. The first exact solution of the discrete-time nonlinear difference equation of SDM without any linearisation approximation and any assumption about the quantisation noise was given by Gray [10] in 1987. He analysed the simplest discrete time model of SDM ( the first order ) with dc input and developed two basic properties:

1) the behaviour of the sigma-delta quantiser and its relation to uniform quantisation;

2) the rate-distortion tradeoffs between the oversampling ratio and the average mean-squared quantisation error.

Furthermore, Gray and his colleagues carried out the spectral analyses of quantisation noise in single-loop sigma-delta modulation with dc and sinusoidal inputs without assuming independent white noise [11][12]. The major conclusion is that the quantiser noise is definitely not white. In fact, it has long been known in practice that single-loop sigma-delta modulator produces spectral noise spikes. It has been shown that the frequency location and weight of the spectral spikes depend in a complex way on the system input [11].

Research activities in this area have become increasingly popular in recent years, particularly due to the proliferation of digital audio systems. Higher order systems such as higher order single stage [8] and multi-stage of cascaded lower order ones (MASH) [13][14] have been proposed to improve the performance of the basic first order system. A single stage structure is realised by combining a higher order noise shaping filter with a single quantiser. This structure was subject to instability if only a one-bit quantiser is used. A stable higher order topology was presented by Chao et. al.[15], where the loop stability is determined primarily by the feed-forward coefficients, while feedback coefficients are added to optimise the signal-to-noise ratio. MASH structure consists of several cascaded lower order (usually first or second order) SDM. Each stage contains one quantiser. This structure avoids the stability problem but is sensitive to component mismatch between individual stages.

The major application of sigma-delta modulation is by far in A/D and D/A conversion with much work directed towards VLSI implementations. A 16-bit performance can be achieved by only using 1-μm CMOS technology and a single 5-V power supply [16]. The VLSI implementation of as high as fifth order single stage SDM has been reported [17]. Currently oversampled A/D and D/A converters can achieve a resolution in the range of 16-20 bits. Recently new applications of SDM have been reported such as FIR filter implementations [18] and waiting time jitter reduction in communication systems [19].

The three major factors in an oversampled sigma-delta modulator are: oversampling ratio, loop filter, and quantiser. Many research simulations and experiments have been carried out with different oversampling ratio, and different combinations of loop filter coefficients to obtain the optimal signal-to-noise ratio and

ensure stability. The main existing methods for designing the loop filter in a SDM are based on the traditional filter design, with which the observed nonlinear phenomena of the system, sometimes come into conflict. Some researchers tried to compromise the conflict by using linear analysis only to determine a starting point of the design. Further design refinements and verification have to be accomplished through computer simulations and breadboard circuit implementation [15]. Therefore, the question arises: is there any design method which does not depend on the concept of linear systems? An optimisation method proposed in this thesis addresses this question. Some more analyses such as the relationship between the quantisation level and the maximum input magnitude are also carried out in order to obtain optimal SDM systems. Usually, once being designed, the quantisation level, the coefficients of the loop filter will be fixed. As we know, adaptation techniques have been successfully used in many systems like digital coding, echo cancellation systems etc. to match changes in operating conditions. No paper has been found to discuss the application of adaptation techniques for SDMs in recent years, although two papers appeared in the early 1970's for simple SDM systems.[45][46] In this thesis, we attempt to adapt SDM systems either in terms of quantisation level or filter coefficients to improve the dynamic range of the systems.

## 1.4 Performance targets for high-quality digital audio systems

Two major factors which decide the performance of audio systems are frequency bandwidth and dynamic range. The first tells how much information in the frequency domain a system will include, which indicates for an analogue-to-digital converting system the minimum speed the clock should have to sample the analogue signal in the

time domain. The second tells the accuracy of magnitude a digital system can represent.

## Frequency bandwidth

For the normal listener, the ear is most sensitive to frequencies between 2 kHz and 5 kHz at the threshold of hearing. The sensitivity drops for frequencies above and below this region such that by 200 Hz and 15 kHz it is approximately 20 dB lower [20]. Some experiments reveal that even highly trained listeners cannot discriminate between conditions of 16- and 20-kHz low-pass cutoff frequencies on programme material containing considerable energy at and above 20 kHz [21][22]. This result supports studies done as early as 1931, which established that a band 40 Hz to 15 kHz is sufficient to reproduce music without an audible change in the reproduction.[35] Nevertheless, recent work examining the low-frequency limits for reproduction suggests that the presence of frequencies below the cutoff of the audible range (20 Hz) can contribute to a more life-like sound quality [23]. In addition, detection of a 20-kHz pure tone is possible for some people at high levels which are greater than 80 dB sound pressure level (SPL) [21]. Thus, for an ideal audio system, a generous choice of bandwidth might be 0 Hz to 20 kHz and an acceptable bandspread would be 20 Hz to 15 kHz. The sampling frequencies must be at least twice higher than the above bandwidth. In digital audio recording systems, the sampling rate is 48 kHz whereas in compact disk playing system, 44.1 kHz sampling frequency is used.

## Dynamic range

The ear's effective dynamic range is 100 dB or more [20]. However, for subjectively noise-free reproduction of music in a quiet environment it has been suggested that approximately 118 dB of dynamic range are required to provide a

listener with the maximum peak SPL and undetectable noise spectrum [24]. Therefore, an ideal system must provide for *playback* somewhere from 100 dB to nearly 120 dB of dynamic range to fulfil the dual requirements of capturing the range of programme material dynamics and matching the ear's range. In uniform quantisation systems, the dynamic range is approximately equal to the maximum signal-to-noise ratio (SNR). However, in nonuniform systems, the dynamic range can be much greater than the maximum SNR. Both compact disk and digital audio tape recording systems use 16-bit uniform quantisation which has 98-dB dynamic range.

## 1.5 Overview of the thesis

The structure of this thesis is depicted in Fig. 1-4, which shows roughly the relations among the chapters.

The first part of this thesis, Chapter 2, is devoted to the basic principle of oversampled sigma-delta modulator (SDM), which is the foundation for the rest of the thesis. The general cases of B bit, nth order SDM system have been described. Up to now, the most common analyses are carried out in the frequency-domain and based on a linear model of the system. This model is not accurate enough for one bit SDM. Although it can still roughly describe the characteristic of the system, sometimes it will be misleading.

The second part of the thesis, Chapter 3, is devoted to the design of stable one-bit SDM. The major parts of a SDM are the loop filter and the quantiser. For the loop filter, the design efforts should concentrate on both signal-to-noise ratio (SNR)

and stability. A trade-off between SNR and stability has to be considered. A new method based on optimisation is introduced to design the loop filter rather than using the traditional filter design methods. A group of optimal (or sub-optimal) coefficients is determined by maximising the signal-to-noise ratio using computer simulations. This will overcome the conflicts occurred in the traditional methods. The concept of the equivalent quantiser is used to investigate the optimal quantisation level with respect to the maximum input level. This chapter is also the base of Chapters 4, 5, and 6.

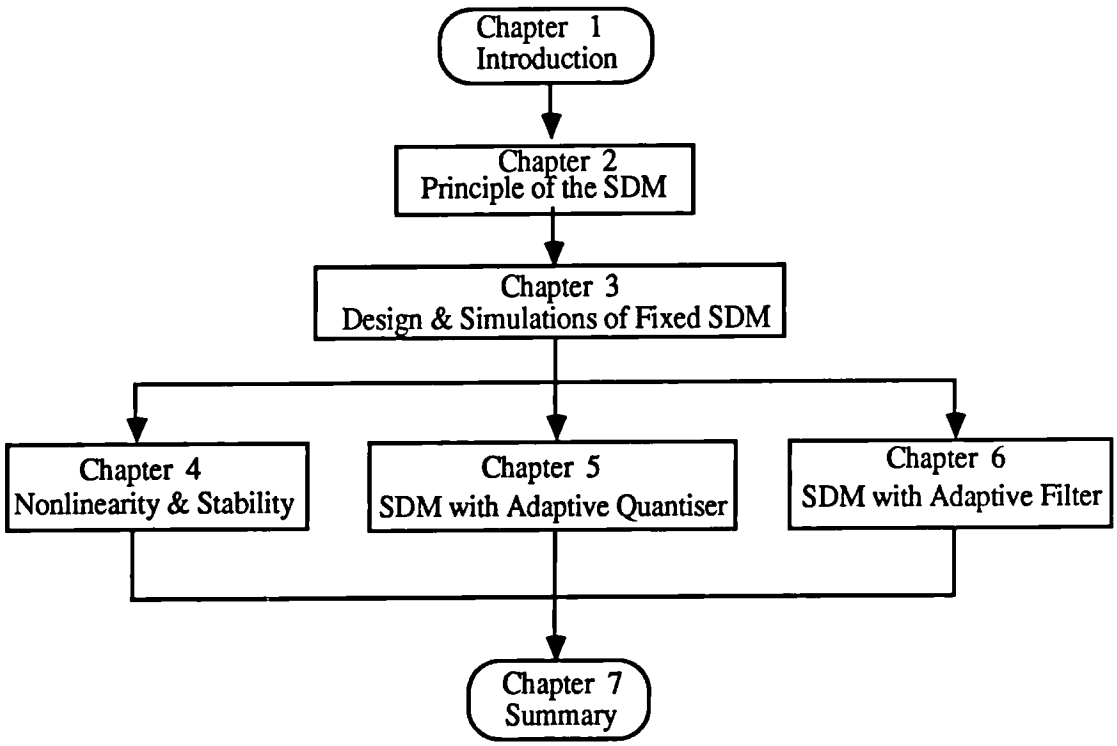



Fig. 1-4 Flowchart of this thesis

In Chapter 4, we discuss the nonlinearity and hence the stability problem of the SDM. The phenomena of the system nonlinearity are demonstrated to show that the linear model of the SDM sometimes will be misleading. Limit cycles are unique features of nonlinear systems. The SDM systems often produce the limit cycles whose average values are used to approach the input. However, sometimes the limit cycles could be very harmful so as to damage the SNR of the system. A method of calculating the period of the limit cycle of the first order SDM is presented when the input is dc. The clippers are designed to prevent the overload of the quantiser.

In Chapter 5, the idea of adaptive quantisation is presented in order to increase the dynamic range of the SDM systems. Although the initial idea appeared in the early 1970's, since then very little research has been carried out. The concept of an equivalent quantiser described in Chapter 3 is used to analyse and design the adaptive quantiser. The feedback digital logic is chosen to estimate the maximum magnitude of the input, upon which the quantisation level is adapted. In the cases of music signals, signals may vary in magnitude very fast during one period of time and relatively slow during another period of time. Therefore, the adaptation speed is crucial. A fast attack time but a slow release time are established to avoid severe overload distortion. When using the adaptive SDM in an A/D or D/A conversion system, the oversampling ratio or the order of the loop filter can be reduced while maintaining the same SNR for small signals at the expense of an increase in noise for the rarely occurring large signals. Many computer simulations are carried out for both sinusoidal and music signals.

Chapter 6 is devoted to adaptive loop filters for sigma-delta modulation. Because of the difficulty of maintaining the stability of high order SDMs, no research results have been published on SDMs with adaptive filters. This chapter describes attempts in adapting some of the coefficients to slightly increase the dynamic range.

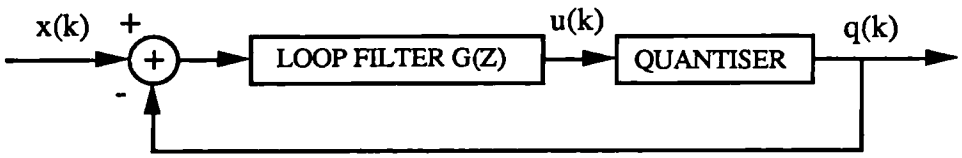
Finally, we summarise in Chapter 7 the major results of this thesis and discuss open problems and future directions in the research of oversampled sigma-delta modulation.

# 2

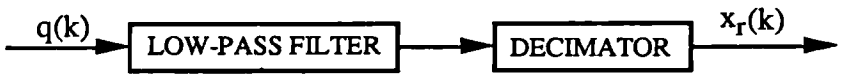
# FUNDAMENTALS OF OVERSAMPLED SIGMA-DELTA MODULATION

## 2.1 Introduction

The basic block diagrams of a digital sigma-delta modulator and a demodulator are depicted in Fig. 2-1. The input x(k) is a signal sampled at a frequency which is much




(a) Modulator




(b) Demodulator

Fig. 2-1 Basic block diagram of sigma-delta modulation

higher than the usual Nyquist rate. This signal combined with the feedback signal then is fed into the loop filter which is basically an integrator, or a cascade of several integrators, the output of which, u(k), will be sent to the quantiser. The quantised signal q(k) will be fed back to the input of the loop filter and simultaneously sent to the demodulator. The loop filter plays a role in reshaping quantisation noise in the frequency domain. The demodulator consists of a low-pass filter and a decimator. The low-pass filter will remove the noise which is outside the signal band. The filter output is then decimated, i.e., has its sampling rate reduced, to obtain the reconstructed in-band signal $x_r(k)$. It is assumed that the wanted signal only occupies the frequency band from 0 to $f_b/2$. If the oversampling ratio is N, the sampling rate $f_s=Nf_b$. Therefore, x(k) is sampled at $f_s$, whereas $x_r(k)$ is sampled at $f_b$.

In this chapter, the principle of sigma-delta modulation (SDM) will be described in detail. Analyses cover from the frequency domain to the time domain, from the modulator to the demodulator, from the oversampling to the noise shaping techniques, from the SDM structure to the alternative one: noise shaper. The two parts of the demodulation: low-pass filter and decimator are also described.

## 2.2 From DM to SDM

Sigma-delta modulation (SDM) was first proposed by Inose et. al.. The initial idea was to pre-emphasise the low frequencies in the input by simply integrating the input prior to delta modulator (DM). This is due to the fact that the output of the DM carries the information which is the differentiation of the input signal. This means it is incapable of transmitting dc component. To compensate for the inevitable differentiation of the

input signal, an integration process is added at the input of the original DM as is shown in Fig. 2-2(a). Because of the added integrator in the modulator, a differentiator is needed in the demodulator as is shown in Fig. 2-2(b). For the realisation of this original configuration, the two integrators can be combined and replaced by an integrator in the forward path of the loop as is shown in Fig. 2-3(a), and the demodulator is simplified to Fig. 2-3(b) which only contains a low-pass filter.

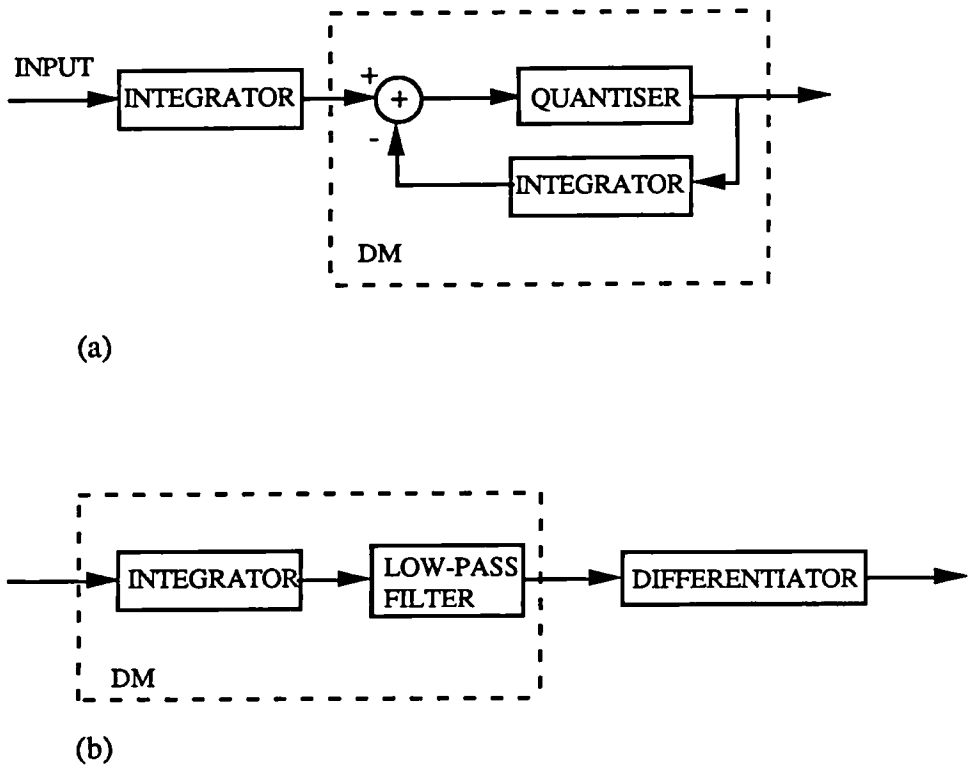
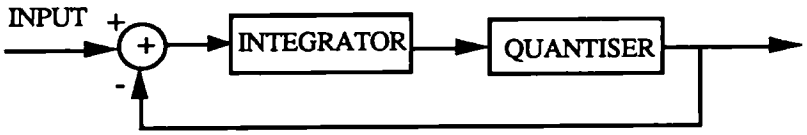



Fig. 2-2 From DM to SDM: (a) adding an integrator at the input of the delta modulator (b) adding a differentiator at the output of the delta demodulator

The integrator inside the loop in Fig. 2-3(a) can be replaced by a general filter which may consist of several cascaded integrators. And the low-pass filter in Fig.2-3(b) can be combined with a decimator to reduce the sampling rate. These will lead to the general block diagram of a SDM system which has been shown in Fig. 2-1.




(a)


(b)

Fig. 2-3 Simplified version of Fig. 2-2: (a) modulator; (b) demodulator

There are intuitive reasons for preferring oversampled SDM to oversampled DM for the modern ADC. For example, there is no error accumulation in the sigma-delta demodulator because there is no feedback loop in the demodulator. Thus, the system is less sensitive to channel errors. Another important difference is that the overload characteristic associated with SDM is independent of the frequency of the input signal while in DM there exists a relationship between input signal frequency and the overload characteristic [25]. In other words, the maximum signal-to-noise ratio (SNR) is independent of the frequency of the input signal in SDM. Furthermore, arguments based on linearised models suggest that the spectral characteristics of the quantisation error are better behaved for SDM than for DM [10].

## 2.3 Oversampling for relaxing pre- and post-filters

An analogue signal is said to be "oversampled" if it is sampled at a rate above (often far above) its Nyquist rate. There is no fundamental (mathematical) reason why a signal-acquisition system needs oversampling. The motivations for doing so derive not from the basic blocks of A/D, D/A converters, but rather from the technology that implements these blocks with finite components, tolerances, and costs. Often, this is an intermediate step in manipulating a signal to be represented ultimately in Nyquist sampled form.

With the current state-of-the-art in analogue circuits, a practical anti-aliasing or anti-imaging filter in conventional (non-oversampling) circuitry can be easily the most expensive element in the signal-acquisition chain of Fig. 1-1. The anti-aliasing filter must pass frequencies below $f_b/2$ which is the highest frequency of the signal band, and suppress frequencies above it . The width of the transition band available around $f_b/2$ in turn constrains the number and stability of time constants (poles and zeros) necessary to realise this analogue filter [26]. The lack of stable precise continuous-time time constants (such as RC products) in standard monolithic fabrication processes implies the need for expensive ( trimmed or discrete component) technology for such an anti-aliasing filter.

Oversampling resolves this problem by sampling initially at an elevated rate $f_s=Nf_b$ when the final sampling rate desired is still $f_b$. An analogue anti-aliasing filter is still necessary but only for anti-alias protection against the high initial sampling rate $Nf_b$. The large difference between the desired signal bandwidth and the new

anti-aliasing cutoff frequency $Nf_b/2$ means that the available transition bandwidth for the filter is now many times the passband width. It permits the analogue filter's magnitude response to roll off gradually and makes it much easier to realise the anti-aliasing filter with imprecise analogue circuitry. Fig. 2-4 illustrates the principle of oversampling for relaxing the pre- and post filters.

In order to accommodate the same final sampling rate $f_b$ as before, the oversampled signal must be further filtered to suppress frequencies above $f_b/2$, but this further filtering can occur digitally after the signal has been quantised.

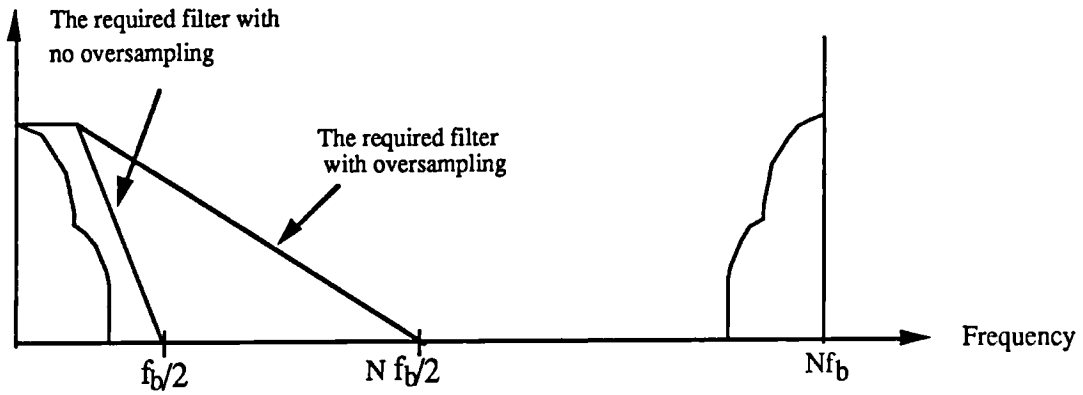



Fig. 2-4 Oversampling for relaxing the filter constraints

## 2.4 Oversampling for resolution enhancement

If an ideal conventional ( sample-by-sample, uniform ) B bit linear A/D converter operates on an input sequence x(k) with an amplitude larger than the least-significant-bit weight and less than the overload level of the quantiser, then the quantisation error introduced by this converter will tend to be input-independent, white noise ( spectrally flat ). This observation motivates the common practice of modelling such an A/D converter as a source of additive white random quantisation noise [27][28][29].

Consider an input x with amplitudes in the range

$$x \in (-X_{max}, X_{max})$$

and a uniform quantiser with a step size

$$\Delta = 2\, X_{max}/2^B \tag{2-1}$$

quantisation error e will have values in the range

$$-\Delta/2 \le e \le \Delta/2$$

If $\Delta$ is sufficiently small, it is reasonable to assume that they are uniform in the above range with a probability density function

$$p(e) = \begin{cases} 1/\Delta, & |e| \le \Delta/2 \\ 0, & \text{otherwise} \end{cases}$$

Thus the variance of the quantisation error is

$$\sigma_e^2 = E[E^2] = \int_{-\infty}^{\infty} e^2\, p(e)\, de = \int_{-\Delta/2}^{\Delta/2} \frac{e^2}{\Delta}\, de = \frac{\Delta^2}{12}$$

where E[ ] denotes expectation, and E inside the bracket represents the random variable of quantisation noise. Using (2-1)

$$\sigma_e^2 = \frac{1}{3} X_{max}^2 \, 2^{-2B}$$

For a sinusoidal input with an amplitude of X, the variance of the signal is

$$\sigma_x^2 = \left(\frac{X}{\sqrt{2}}\right)^2 = \frac{X^2}{2}$$

so that the signal-to-noise ratio (SNR) is

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \left(\frac{3}{2}\right) 2^{2B} \left[\frac{X}{X_{max}}\right]^2$$

where X is the (zero-to-peak) sinusoid amplitude and $X_{max}$ is the maximum input amplitude of the A/D converter. This SNR reaches a maximum value when the sinusoid amplitude just fills (saturates) the converter's input range ($X=X_{max}$), so that as a power ratio,

$$SNR_{max} = (3/2)\, 2^{2B}$$

or in decibel form

$$SNR_{max}\ (dB) \approx (6.02)B + 1.76 \qquad (2\text{-}2)$$

Signal waveforms other than sinusoids will yield an additive term different from 1.76 dB if their peak-to-RMS ratio differs from that of a sinusoid. Fig. 2-5 shows a typical SNR curve versus signal amplitude for a linear (uniform) A/D converter. When the input amplitude is larger than $X_{max}$, the overload distortion will occur.

The demands on the quantiser for a given ultimate signal resolution can be relaxed by oversampling and then low-pass filtering. As was mentioned before, the quantisation noise will tend to be wideband, white noise. The spectrum of Fig. 2-6(a), with no oversampling, shows quantisation noise occupying the same frequency range as the input signal of interest. In contrast, Fig. 2-6(b) shows quantisation occurring at

an oversampled rate $Nf_b$, where again $f_b/2$ is the analogue band width of interest. If the quantiser is a simple B-bit A/D converter, its quantisation-noise power does not depend on its sampling rate, but a higher sampling rate will spread this power over a wider range of frequencies. Subsequently filtering out frequencies above $f_b/2$ with a digital low-pass filter will reduce the quantisation-noise power, effectively increasing the resolution of the quantiser. The decimator will drop the sampling rate to $f_b$ once frequencies above $f_b/2$ are removed.

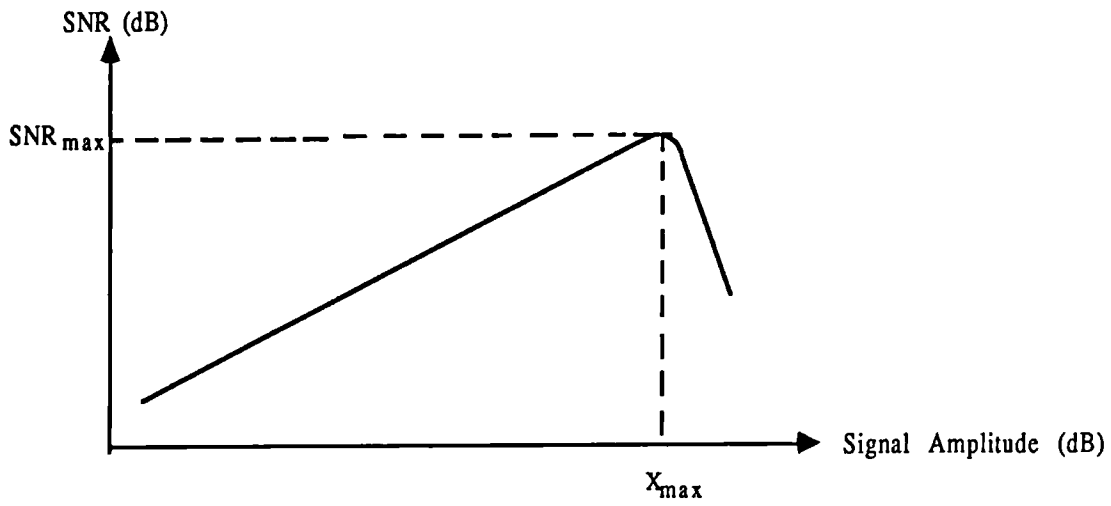



Fig. 2-5 Representative curve of SNR versus signal amplitude for a linear A/D converter

With the oversampling and decimation factor of N in Fig. 2-6, and white quantisation noise from the quantiser, an ideal low-pass filter will reduce the quantisation noise power by a factor of N while leaving the signal power unaffected.

This means

SNR Enhancement = N

Thus for a B-bit linear converter, the peak SNR in dB for a sinusoidal signal becomes

$$SNR_{max}\ (dB) \approx (6.02)B + 1.76 + 10\log_{10}N \qquad (2\text{-}3)$$

It is also convenient to write $N=2^L$, so that L is the number of octaves of oversampling. Then (2-3) can be rearranged to

$$SNR_{max}\ (dB) \approx 6.02\ (\ B + 0.5L\ ) + 1.76 \qquad (2\text{-}4)$$

Equation (2-4) shows directly that the oversampling A/D converter yields baseband SNR equivalent to that of a non-oversampling converter with a higher number of bits, and in this case the tradeoff is 0.5 bits per octave of oversampling.

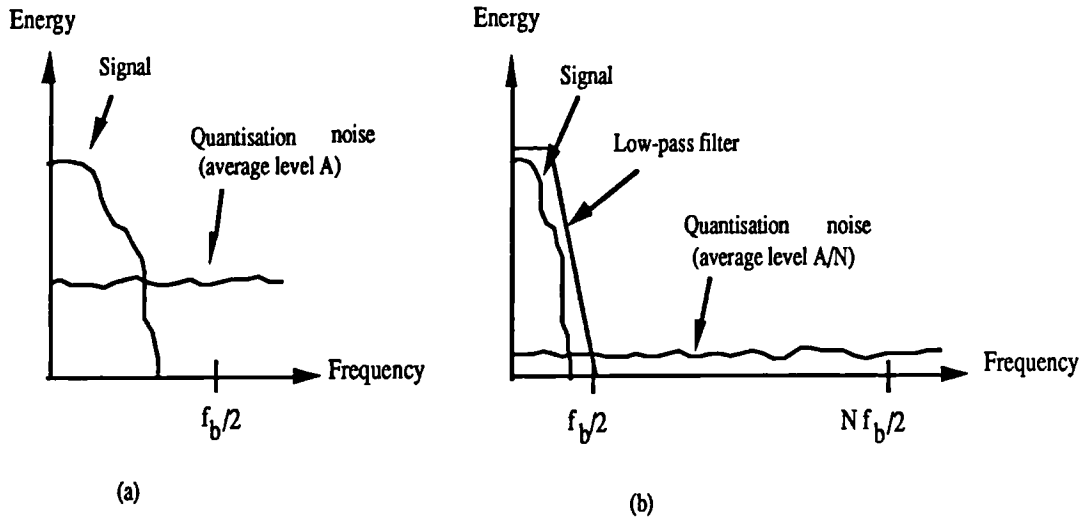



Fig. 2-6 Oversampling for resolution enhancement: (a) spectrum with no oversampling (b) spectrum with oversampling

## 2.5 Noise shaping function

From the previous section 2.4, intuitively, the oversampling and low-pass filtering process is just a time-domain averaging by which coarse digital output codes of a quantiser are "interpolated" in level to yield a finer quantisation. However, efficient arrangement of the whole system by carefully designing the loop filter G(z) in Fig. 2-1, to spectrally shape the quantiser's error, yields more-than-intuitive resolution enhancements through oversampling. In this section we will show how the loop filter G(z) plays the role of spectrally reshaping the quantisation noise. From Fig. 2-1(a), the system can be described as follows:

$$[\, X(z) - Q(z) \,]\, G(z) = U(z) \tag{2-5}$$

Let

$$u(k) = q(k) - e(k) \tag{2-6}$$

where e(k) is the quantisation noise, so that

$$U(z) = Q(z) - E(z) \tag{2-7}$$

Combining (2-5) and (2-7),

$$Q(z) = \frac{G(z)}{1+ G(z)} X(z) + \frac{1}{1+ G(z)} E(z) \tag{2-8}$$

This can be generalised to

$$Q(z) = F_X(z)X(z) + F_E(z)E(z) \tag{2-9}$$

where $F_X(z)$ and $F_E(z)$ are the signal and noise transfer functions respectively

$$F_X(z) = \frac{G(z)}{1+ G(z)} \quad , \qquad F_E(z) = \frac{1}{1+ G(z)} \tag{2-10}$$

Supposing that the analogue input signal has some total power $\sigma_x^2$ distributed in frequency according to a power spectral density (PSD) $S_X(\lambda)$, while the source of

quantisation noise in the system has power $\sigma_e^2$ and power spectral density $S_E(\lambda)$, normally, but not necessarily, white. Here, $\lambda$ is the normalised frequency variable for discrete time, taking the range 0 to $2\pi$, where $\lambda = 2\pi$ corresponds to a physical (Hertz) frequency equal to the current sampling rate.

In general the system described by equation (2-9) may frequency-filter the signal input as well as the quantisation noise; both must be considered. From (2-9), at the modulator output q(k), the signal-component PSD, $S_{Qs}(\lambda)$, and the noise-component PSD, $S_{Qn}(\lambda)$, are

$$S_{Qs}(\lambda) = |F_X(e^{j\lambda})|^2 S_X(\lambda) \qquad S_{Qn}(\lambda) = |F_E(e^{j\lambda})|^2 S_E(\lambda) \qquad (2\text{-}11)$$

Now the total signal power and noise power in q(k) over the baseband of interest are respectively (power in 1 Ω),

$$\sigma_{bs}^2 = \frac{1}{\pi}\int_0^{\pi/N} S_{Qs}(\lambda)\, d\lambda = \frac{1}{\pi}\int_0^{\pi/N} |F_X(e^{j\lambda})|^2 S_X(\lambda)\, d\lambda \qquad (2\text{-}12a)$$

$$\sigma_{bn}^2 = \frac{1}{\pi}\int_0^{\pi/N} S_{Qn}(\lambda)\, d\lambda = \frac{1}{\pi}\int_0^{\pi/N} |F_E(e^{j\lambda})|^2 S_E(\lambda)\, d\lambda \qquad (2\text{-}12b)$$

The ratio of these two baseband powers is the output SNR of the oversampling A/D converter. Its most general form is

$$SNR = \frac{\sigma_{bs}^2}{\sigma_{bn}^2} = \frac{\int_0^{\pi/N} |F_X(e^{j\lambda})|^2 S_X(\lambda)\, d\lambda}{\int_0^{\pi/N} |F_E(e^{j\lambda})|^2 S_E(\lambda)\, d\lambda} \qquad (2\text{-}13)$$

With the topology in Fig. 2-1, exhibiting the signal and noise transfer functions of (2-10), and with white quantisation noise from the internal A/D converter ( making $S_E(\lambda)$ a constant ), (2-13) becomes

$$SNR = \frac{\int_0^{\pi/N} \left| \frac{G(e^{j\lambda})}{1+G(e^{j\lambda})} \right|^2 S_X(\lambda)\, d\lambda}{\sigma_e^2 \int_0^{\pi/N} \frac{1}{|1+G(e^{j\lambda})|^2}\, d\lambda} \qquad (2\text{-}14)$$

where $\sigma_e^2$ is the power of quantisation noise e(k) over the entire band $(0, 2\pi)$.

To further evaluate the SNR requires knowledge of the specific loop-filter function G(z). For the nth order sigma-delta modulator, the most commonly used function G(z) is [30]

$$G(z) = \frac{1}{(1 - z^{-1})^n} - 1 \qquad (2\text{-}15)$$

Therefore, equation (2-8) becomes

$$Q(z) = [\, 1 - (1-z^{-1})^n \,]\, X(z) + (1-z^{-1})^n\, E(z)$$

that is,

$$F_X(z) = 1 - (1-z^{-1})^n \quad , \qquad F_E(z) = (1-z^{-1})^n \qquad (2\text{-}16)$$

From equation (2-16), it can be seen that when frequency f is much less than $f_s$, $|F_X(e^{j2\pi f})|$ is approximately equal to one. This is the case for a baseband signal in a highly oversampled system. It can also be seen that the quantisation noise has been shaped. $F_E(z)$ behaves like a high-pass filter or differentiator. It moves most of the noise from the low frequency portion of the available bandwidth to the high frequency portion so that the noise in the signal band can be reduced. The frequency-response

magnitude of the noise transfer function in (2-16), at normalised frequency $\lambda$, is

$$| F_E(e^{j\lambda}) | = [ ( 2 - 2\cos\lambda )^{1/2} ]^n = [ 2\sin(\lambda/2) ]^n \qquad (2\text{-}17)$$

Fig. 2-7 shows the curves of (2-17) when n=1, 2, and 3.

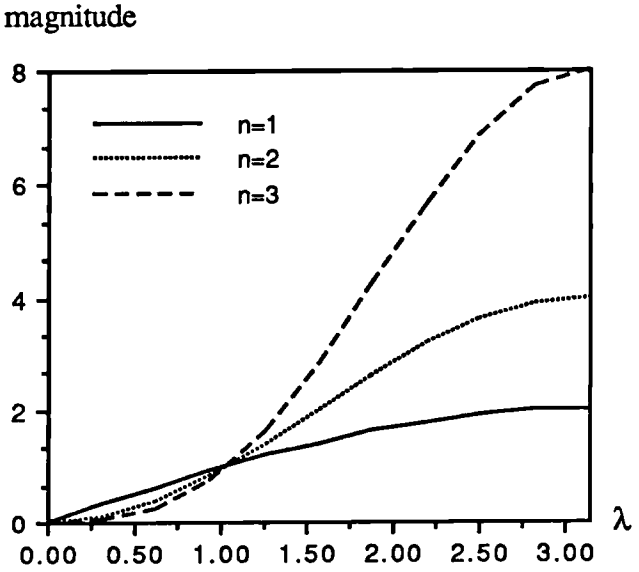



Fig. 2-7 Magnitude spectra of noise shaping functions

$|F_X(e^{j2\pi f})|$ is approximately equal to one in the baseband so that the numerator of (2-14) simplifies to

$$\int_0^{\pi/N} S_X(\lambda)\, d\lambda = \pi\, \sigma_x^2$$

where $\sigma_x^2$ is just the original analogue-input-signal power ( this signal is in the baseband and therefore $S_X(\lambda)$ is non-zero only between $\lambda$=0 and $\lambda$=$\pi$/N ). From

(2-17) the denominator integral in (2-14) is

$$\sigma_e^2 \int_0^{\pi/N} 4^n \sin^{2n}(\frac{\lambda}{2})\, d\lambda$$

So the SNR at output is

$$SNR = \frac{\sigma_{bs}^2}{\sigma_{bn}^2} = \left[\frac{\sigma_x^2}{\sigma_e^2}\right] \frac{\pi}{4^n \int_0^{\pi/N} \sin^{2n}(\frac{\lambda}{2})\, d\lambda} \qquad (2\text{-}18)$$

The first factor, $\sigma_x^2/\sigma_e^2$, is just the SNR that would result if the same internal A/D converter operated directly on the signal input x(k) with no oversampling and noise-shaping. Thus the remaining factor in the righthand side of (2-18) is a net SNR enhancement attributable to the oversampling-decimating process with the topology of Fig. 2-1 and the particular loop filter of (2-15). Therefore, in power-ratio terms, the SNR enhancement for the nth order system with N times oversampling rate is

$$\text{SNR Enhancement} = \frac{\pi}{4^n \int_0^{\pi/N} \sin^{2n}(\frac{\lambda}{2})\, d\lambda} \qquad (2\text{-}19)$$

Table 2-1 gives the final results of (2-19) when n=1, 2, and 3. For large oversampling factors $N>>\pi$, (2-19) can be approximated using a Taylor expansion of the sine function as follows

$$\text{SNR Enhancement} = (2n+1)N^{2n+1}/(\pi^{2n})$$

Again, let $N=2^L$ so that the SNR Enhancement in decibel form is

$$\text{SNR Enhancement (dB)} = 6.02\,(n + 0.5)\,L + 10\log_{10}\left(\frac{2n+1}{\pi^{2n}}\right) \qquad (2\text{-}20)$$

The specific case of a maximum-amplitude sinusoid input signal, which in (2-2) yielded a maximum SNR of 6.02B + 1.76 dB from a B-bit linear A/D in the absence of oversampling, will now achieve a maximum SNR of

$$SNR_{max}\ (dB) \approx 6.02\ [\ B + (n+0.5)L\ ] + 10\log_{10}(2n+1) - 9.943n + 1.76 \qquad (2\text{-}21)$$

Thus, for the first, second, third, and fourth order systems, the $SNR_{max}$ in dB can be obtained as follows

$$SNR_{max}\ (dB) \approx 6.02\ (\ B + 1.5L\ ) - 3.14 \qquad \text{(1st order)} \qquad (2\text{-}22a)$$

$$SNR_{max}\ (dB) \approx 6.02\ (\ B + 2.5L\ ) - 11.14 \qquad \text{(2nd order)} \qquad (2\text{-}22b)$$

$$SNR_{max}\ (dB) \approx 6.02\ (\ B + 3.5L\ ) - 19.62 \qquad \text{(3rd order)} \qquad (2\text{-}22c)$$

$$SNR_{max}\ (dB) \approx 6.02\ (\ B + 4.5L\ ) - 28.47 \qquad \text{(4th order)} \qquad (2\text{-}22d)$$

Table 2-1 SNR enhancement of the SDM system

| Order of the system | SNR enhancement | Approximation of SNR enhancement (for $N \gg \pi$) |
|---|---|---|
| n=1 | $\dfrac{\pi}{2\ [\ \pi/N - \sin(\pi/N)\ ]}$ | $\dfrac{3N^3}{\pi^2}$ |
| n=2 | $\dfrac{\pi}{[6\pi/N - 8\sin(\pi/N) + \sin(2\pi/N)]}$ | $\dfrac{5N^5}{\pi^4}$ |
| n=3 | $\dfrac{\pi}{[20\pi/N - 30\sin(\pi/N) + 6\sin(2\pi/N) - (2/3)\sin(3\pi/N)]}$ | $\dfrac{7N^7}{\pi^6}$ |

From (2-21) it can be seen that increasing orders of high-pass shaping in the quantisation noise will tend to increase the resolution payoff. Note, the noise-shaping function described by (2-16) is probably the simplest function in the aspect of reducing the noise power in the low-frequency portion. However, it also boosts its total power. This results in the fixed losses in SNR in (2-22) with the effect becoming more pronounced as the order of the noise-shaping loop grows. In some audio applications, the high power gain of the noise shaping function at high frequencies is unwanted because of the fact that some systems such as PWM digital power amplifier often tend to give high-frequency intermodulation. Also, high power gain may cause the system to be unstable for the higher order, very low bit systems. Thus, a loop filter which differs from (2-15) has to be designed, especially for higher order systems, which will be discussed in detail in the following chapters.

## 2.6 Sigma-delta modulator and noise shaper

An alternative structure is called a noise shaper; its basic block diagram is shown in Fig. 2-8. It can be derived from Fig. 2-8 that

$$Q(z) = X(z) + ( 1-H(z) ) E(z)$$

Usually, for the nth order noise shaper

$$H(z) = 1 - (1-z^{-1})^n$$

so that

$$Q(z) = X(z) + (1-z^{-1})^n E(z)$$

Therefore,

$$F_X(z) = 1, \quad F_E(z) = (1-z^{-1})^n \tag{2-23}$$

Comparing equations (2-16) and (2-23), it can be seen that the noise is reshaped in the

same manner. However these two systems are slightly different with respect to the input signal, in that the in-band gain is identically one for the noise shaper but only approximately so for a SDM. The relationships between G(z) and H(z) are

$$H(z) = \frac{G(z)}{1 + G(z)} \quad \text{or} \quad G(z) = \frac{H(z)}{1 - H(z)}$$

Fig. 2-9 shows the block diagram of the sigma-delta modulator that results from the noise shaper [31]. Considering that H(z) closely approximates 1 in the signal band, i.e.,supposing that $F_{x(\text{sigma-delta})}=F_{x(\text{noise shaper})}=1$, the effect of 1/H(z) in Fig. 2-9 can be ignored. It can be seen that there are no big differences between sigma-delta modulators and noise shapers. The structures are different, but describe the same thing.

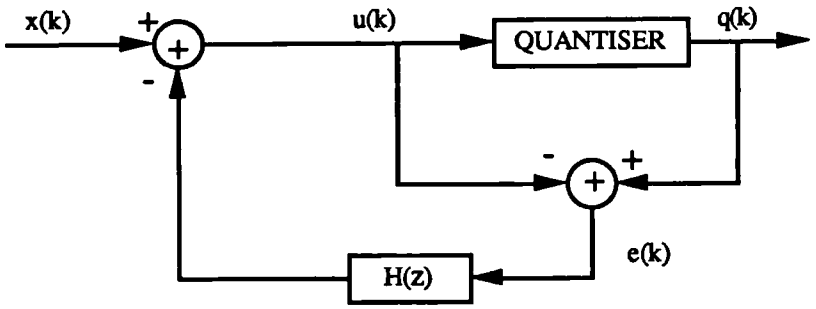



Fig. 2-8 Basic diagram of noise shaper

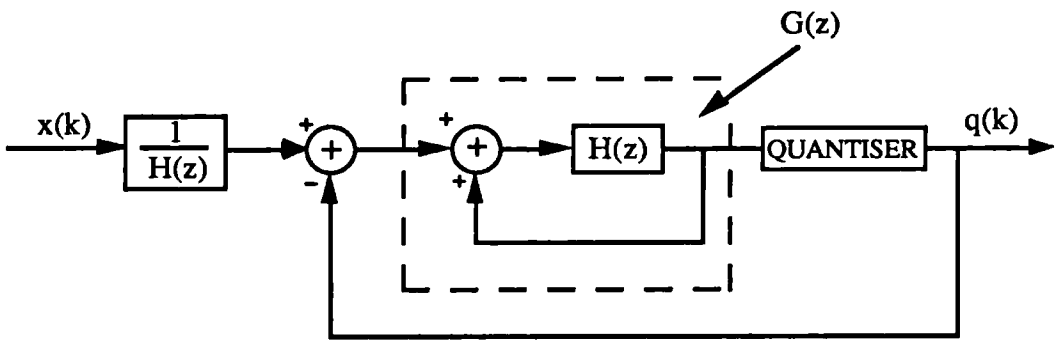



Fig. 2-9 Sigma-delta modulator that results from the noise shaper

## 2.7 One-bit sigma-delta modulators

One-bit oversampled A/D converters differ from multi-bit converters in many ways:

1) It is much easier to realise "good" one-bit internal A/D and D/A elements than multi-bit elements in most solid-state technologies. In practice, a converter having more than a single bit tends to have differential nonlinearities due to mismatch in its current sources, etc. A one-bit converter contains only two reference values: one and zero. Any errors give only a gain and/or an offset error but not a linearity error. As a result, a one-bit conversion system has excellent differential linearity. However, in multi-bit cases, dynamic element matching, laser trimming, and current calibration are needed to achieve high accuracy.

2) When the input to the low-pass filter in the demodulator is only one bit wide, it can greatly simplify the digital filter arithmetic and hence the implementation of the low-pass filter [32]. In particular, a finite impulse response (FIR) first stage in the comb filter requires no full multiplications, since FIR arithmetic can be arranged with an input sample as a factor in every product.

3) Other aspects being equal, one-bit A/D converters require a higher oversampling ratio N than multi-bit versions, since essentially all of their resolution arises from oversampling.

4) Because of the coarse quantisation, nonlinearity of the system must be considered (note: this is different from the differential nonlinearity error of the quantiser circuit ). The model in (2-6) is not adequate enough. Analyses based on that model can still predict broad behaviour when B=1. For example, the SNR can be predicted to decrease with the decrease of the input level, oversampling ratio, and the order of the loop filter. However , they can also be misleading in the sense of precise SNR value or stability. A new model needs to be established for more precise analysis.

5) One-bit sigma-delta modulators are usually less stable than multibit converters, especially in the case of high order systems. This makes the design of one-bit high order sigma-delta modulators much more difficult.

## 2.8 Dynamic analysis in the time domain

This section gives the analysis of the sigma-delta modulators from time domain point of view. First, in order to show the dynamic process more clearly the first order model with one-bit quantiser is chosen. Higher order and multi-bit systems have similar properties. Second, a dc input is assumed. This is because [33][10]:

(i) constant inputs simplify analysis; a number of illuminative results have been based on this simplifying assumption;

(ii) the oversampling of the input in practical situations implies that it appears approximately constant to the modulator;

(iii) any modulator designed for dynamic inputs must be able to handle constant inputs as a special case.

The discrete time model of the first order sigma-delta modulator is depicted in Fig. 2-10, where Q is a one-bit quantiser given by

$$Q(u) = \begin{cases} d, & u \geq 0 \\ -d, & u < 0 \end{cases}$$

From Fig. 2-10, it can be seen that

$$u_k = u_{k-1} + v_{k-1} \tag{2-24a}$$

$$v_k = x - q_k \tag{2-24b}$$

where x is a constant input and $q_k$ is either d or -d. Substituting (2-24b) into (2-24a),

it can be obtained that

$$u_k = u_{k-1} + x - q_{k-1} = u_0 + kx - \sum_{i=1}^{k} q_{i-1} \tag{2-25}$$

so that

$$\frac{1}{k}\sum_{i=0}^{k-1} q_i = x - \frac{u_k - u_0}{k} \tag{2-26}$$

Equation (2-26) means that the average of the output over k samples will approach the input x when k approaches infinity ( supposing $u_k$ is finite). If k is finite, then the error will be $(u_k-u_0)/k$.

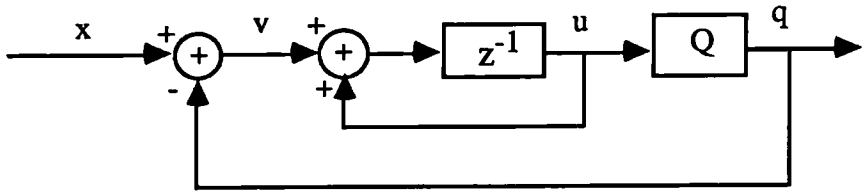



Fig. 2-10 Discrete time model of the first order sigma-delta modulator

According to [34], if the input to a sigma-delta modulator is a dc level which can be expressed as a rational number b/a, when normalised with respect to the quantiser step, the output bit string is periodic, that is, a limit cycle occurs. In this case, the average of the output over a period P will become

$$\frac{1}{P}\sum_{i=0}^{P-1} q_i = x \tag{2-27}$$

This is because $u_P = u_0$. Therefore, if the average is taken over the exact period time, then the error will be zero. Limit cycles can thus be seen as a natural result of

approximating constant inputs using sigma-delta modulators, as evidenced by their prominent position in several papers, including [31] [6]. It is shown that whether or not the binary quantiser output is periodic, its time average should approximate the dc input [10].

We now consider an equivalent quantisation noise model. Define the binary quantiser error as

$$e_k = Q(u_k) - u_k = q_k - u_k$$

The quantiser outputs can be expressed in terms of the binary quantiser error as

$$q_k = u_k + e_k \tag{2-28}$$

Substituting $u_k$ with (2-25), (2-28) becomes

$$q_k = u_{k-1} + x - q_{k-1} + e_k = x + e_k - e_{k-1} \tag{2-29}$$

If the average of both sides of (2-29) is taken, then

$$\frac{1}{k}\sum_{i=1}^{k} q_i = x + \frac{1}{k}\sum_{i=1}^{k} (e_i - e_{i-1}) = x + \frac{e_k - e_0}{k} \tag{2-30}$$

Considering the left-hand side of (2-30) as an equivalent quantiser to x, that is

$$Q_x(x) = \frac{1}{k}\sum_{i=1}^{k} q_i \tag{2-31}$$

then,

$$Q_x(x) - x = \frac{e_k - e_0}{k}$$

so that the quantisation noise is k times smaller. And it also can be seen that the quantisation noise of x is dependent on the initial condition $e_0$. If we choose oversampling ratio to be N, and take average value over N samples, then the error will be roughly N times smaller than the quantisation noise caused by single bit quantiser. It can be proved that the upper bound on the absolute error is inversely proportional to

the oversampling ratio N. Gray [10] has proved that if the initial state $u_0$ is in the range [x-d,x+d], then the state $u_k$ remains in the same range at all future times k. Suppose that the input is within the range [-d, d), and the initial state $u_0$ is within [x-d, x+d), then by using (2-26), where k=N, the following can be obtained

$$x - \frac{1}{N}\sum_{i=0}^{N-1} q_i = \frac{u_N - u_0}{N}$$

so that

$$| x - Q_x(x) | \le \frac{x+d - (x-d)}{N} = \frac{2d}{N} \qquad (2\text{-}32)$$

The above states that the maximum quantisation error for an input within the dynamic range of the quantiser and an initial condition within d of the input is inversely proportional to the oversampling ratio N. Therefore, the quantisation noise decreases as the oversampling ratio increases, which is consistent with the result from frequency domain analysis in Section 2.2.

Supposing that in (2-31) over k samples of q, $k_1$ of them possess the value d and $k_2$ are -d, where $k_1+k_2=k$, we can obtain that

$$Q_x(x) = d(k_1-k_2)/k \qquad (2\text{-}33)$$

The above equation means that the value of $Q_x(x)$ not only relates to the occurring times of the positive and negative samples, but also the quantisation level d.

An example is given in Table 2-2. Assume that $u_0=0$, x=0.5, the quantisation level d=1. It shows that as k becomes larger and larger, the average value of $q_k$ is closer and closer to the dc input. It also shows that at the points k equal to the period

length or an integer multiple of it, the average value is exactly equal to the dc input. The corresponding curves are given in Fig. 2-11.

Table 2-2 An example of SDM's principle with dc input

| $x = 0.5$ | | $u_k = u_{k-1} + x - y_{k-1}$ | | |
|---|---|---|---|---|
| k | $u_k$ | $y_k = Q(u_k)$ | $Q_x(x) = \frac{1}{k}\sum_{i=0}^{k-1} y_i$ | Error=$\lvert x - Q_x(x) \rvert$ |
| 0 | 0.0 | 1 | 1.0000 | 0.5000 |
| 1 | -0.5 | -1 | 0.0000 | 0.5000 |
| 2 | 1.0 | 1 | 0.3333 | 0.1667 |
| 3 | 0.5 | 1 | 0.5000 | 0.0000 |
| 4 | 0.0 | 1 | 0.6000 | 0.1000 |
| 5 | -0.5 | -1 | 0.3333 | 0.1667 |
| 6 | 1.0 | 1 | 0.4286 | 0.0714 |
| 7 | 0.5 | 1 | 0.5000 | 0.0000 |
| 8 | 0.0 | 1 | 0.5556 | 0.0556 |
| 9 | -0.5 | -1 | 0.4000 | 0.1000 |
| 10 | 1.0 | 1 | 0.4545 | 0.0455 |
| 11 | 0.5 | 1 | 0.5000 | 0.0000 |
| 12 | 0.0 | 1 | 0.5385 | 0.0385 |
| 13 | -0.5 | -1 | 0.4286 | 0.0714 |
| 14 | 1.0 | 1 | 0.4667 | 0.0333 |
| 15 | 0.5 | 1 | 0.5000 | 0.0000 |
| . . . | . . . | . . . | . . . | . . . |
| 100 | 0.0 | 1 | 0.5049 | 0.0049 |
| 101 | -0.5 | -1 | 0.4951 | 0.0049 |
| 102 | 0.5 | 1 | 0.4903 | 0.0097 |
| 103 | 1.0 | 1 | 0.5000 | 0.0000 |
| 104 | 0.0 | 1 | 0.5048 | 0.0048 |
| . . . | . . . | . . . | . . . | . . . |

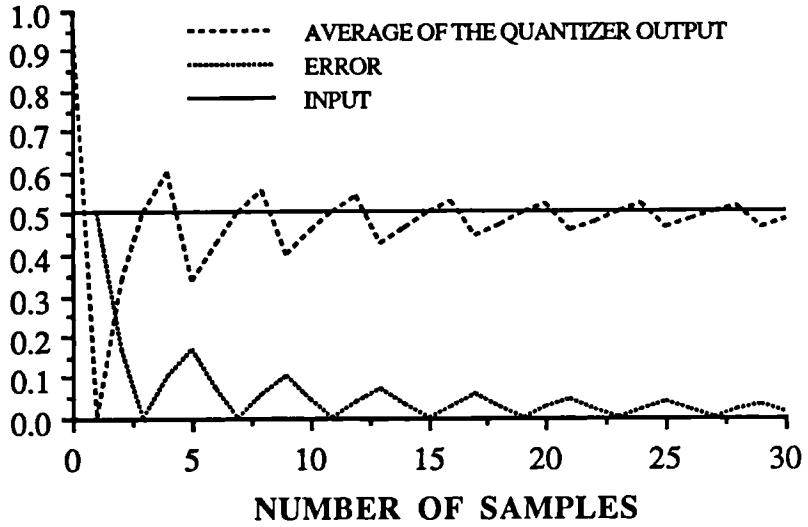



Fig. 2-11 An example of SDM's principle with dc input

The averaging function in (2-27) has the same property as a low-pass filter. Using a low-pass filter different from (2-27) can still achieve the similar result, that is, the low-pass filtered output of the quantiser should approach the dc input provided that the magnitude of the filter is one.

In general, whether the input is dc or not, the sigma-delta modulator tries to minimise the noise by having the quantised values oscillate between levels in such a way that the average of the output $q_k$ approximates the average of the input $x_k$.

## 2.9 Low-pass filters and decimators

If the oversampling ratio is very high ( in most of the cases, higher than 64 ), a very narrow band low-pass filter is needed to remove the out-of-band noise. In general, to design such a filter is quite difficult and the order of the filter has to be very high, which leads to complex implementation.

Usually, a comb filter [9] is used to avoid this problem. Its transfer function is:

$$H(Z) = \left(\frac{1}{N}\frac{1 - Z^{-N}}{1 - Z^{-1}}\right)^{n}$$

$$|H(e^{j\omega})| = \left(\frac{1}{N}\frac{\sin\frac{\omega N}{2}}{\sin\frac{\omega}{2}}\right)^{n} \qquad (2\text{-}34)$$

where N is the oversampling ratio, n is called the stage number of the comb filter. For n=1, the impulse response is

$$h_1(i) = \begin{cases} \frac{1}{N}, & i = 0,1,\ldots,N-1 \\ 0, & \text{otherwise} \end{cases}$$

$h_1(i)$ is a rectangular function, which is an (N-1)th order linear phase finite impulse response filter. Its Fourier transform is:

$$H_1(e^{j\omega}) = \frac{1}{N}\frac{\sin(\frac{\omega N}{2})}{\sin(\frac{\omega}{2})}e^{-j(N-1)\omega/2}$$

The magnitude response of $H_1(e^{j\omega})$ is shown in Fig. 2-12, where N=16. The

relationship between the input q(k) and the output y(k) of the low-pass filter $H_1(z)$ is quite simple:

$$y(k) = \frac{1}{N} \sum_{i=0}^{N-1} q(k-i) \tag{2-35}$$

which can be implemented easily in the feedback form

$$y(k) = y(k-1) + [q(k)-q(k-N)]/N \tag{2-36}$$

For n>1, the system can be considered as a cascade of filters with n=1.

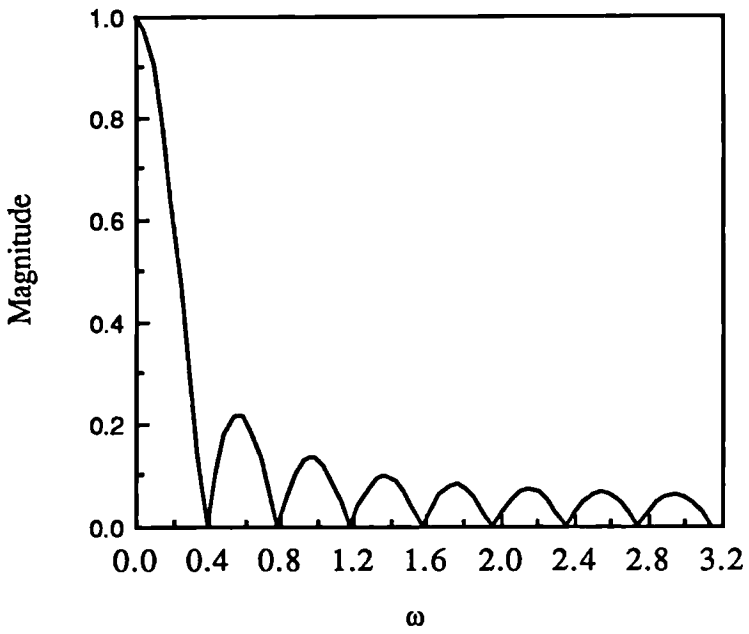



Fig. 2-12 Magnitude response of the 16th order comb filter

In some audio applications, other audio equipment like analogue amplifiers often tend to give high-frequency intermodulation. Therefore the stage number n of the low-pass filter in equation (2-34) should be at least one higher than the order of the loop filter in order to suppress the high-frequency noise sufficiently.[9] For example, if the 3rd

order loop filter is chosen in a sigma-delta modulator, the stage number n in equation (2-34) should be at least 4.

From equations (2-35) and (2-36), it can be seen that the comb filter is easy to design and to implement. It has no multiplications. The transfer function is approximately 1 near the point of $\omega=0$, but as the frequency increases, the gain decreases. At the upper frequency of the signal band ($\omega=\pi/N$), the gain becomes

$$\left|H(e^{j\frac{\pi}{N}})\right| = \left|\frac{1}{N}\frac{\sin\frac{\pi}{2}}{\sin\frac{\pi}{2N}}\right|^n = \left|\frac{1}{N}\frac{1}{\sin\frac{\pi}{2N}}\right|^n \approx \left|\frac{1}{N}\frac{2N}{\pi}\right|^n = \left(\frac{2}{\pi}\right)^n$$

The bigger the value n is, the bigger the attenuation of out-of-band noise, but the more severe the frequency distortion. Even when n=1, an amplitude reduction at the high-frequency end of the signal-band will be $2/\pi$, which is 3.92 dB.

Some compensations have been used for solving this problem [35]. From equation (2-34), it can be seen that when $\omega$ is very small, $\sin(\omega/2)\approx\omega/2$ so that $H(e^{j\omega})$ is like the function $(\sin x/x)^n$. A sinx/x corrector from [35] is shown in Fig. 2-13. Another method can be used to reduce the distortion [36]. The intermediate oversampling ratio R ( R<N ) is chosen such that the frequency response of the comb filter is close to unity in the signal band and that the complexity of the signal-band filter can be kept low. This method is used for the simulations of one-bit sigma-delta modulation in this thesis. The structure is shown in Fig. 2-14, where N=4R in this case. It can be seen that it is much easier to design a half-band filter with a nearly ideal low-pass function.

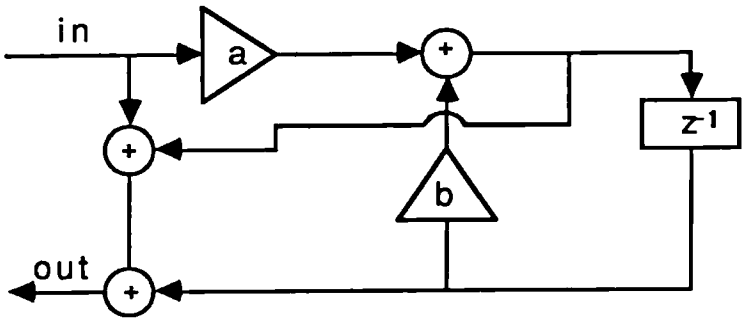



Fig. 2-13 Flow-graph of the sinx/x corrector

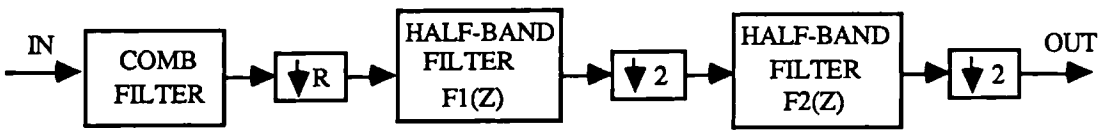



Fig. 2-14 Implementation of low-pass filter and decimator

## 2.10 Summary

In this chapter, the principle of SDM has been described. From the frequency domain angle, the principle of sigma-delta modulation is the spreading of the quantisation noise in a band which is much larger than that of the signal by oversampling and the further reduction of the in-band noise by reshaping the whole band noise. From the time domain angle, the coarse quantisations causes rapid oscillations between levels,

keeping their running average representative of the input. The oversampling technique can also relax the anti-aliasing and anti-imaging filters.

In frequency domain analysis, an additive independent white quantisation noise model is assumed, from which the approximate maximum SNR can be derived. It was seen to be a function of the bit number, the oversampling ratio, and the order of the loop filter. In the next chapter, it can be seen that the maximum SNR functions (2-22) are more precise for the multi-bit case than the one-bit case. For a one-bit SDM, the simulation results will be much worse than the predicted $SNR_{max}$ by (2-22) because of imprecise model of quantisation noise.

Exact analysis in the time-domain can be carried out for the first order one-bit case, in which the upper bound of the absolute error can be derived.

Multi-stage comb filters are often used as the first parts of the low-pass filter because of their simplicity. The number of stages is normally at least one larger than the order of the loop filter in order to suppress the out-of-band noise sufficiently.

# 3

# DESIGN OF A STABLE ONE-BIT SIGMA-DELTA MODULATOR

## 3.1 Introduction

In the previous chapter, the basic principle of sigma-delta modulation systems has been described, which is based on the ideal noise transfer function $(1-z^{-1})^n$. Computer simulations from the author and colleagues [56] have shown that this function can work well for higher order and multi-bit ( usually more than 3 bits ) SDM systems. However, for a one-bit SDM system, when the order n is chosen higher than two, the above transfer function will normally cause the system to be unstable so as to offset the average of the output far from the input value. Therefore, for the noise shaping filter, the problem is how high the order should be, what kind of structure it should have, and how to choose the coefficients so as to suppress the base band noise to the minimum level and to prevent the system from being unstable.

For the quantiser, the problem is how to set up the quantisation level so as to match the maximum possible input. In other words, with a given quantisation level,

what is the maximum input level which does not cause severe distortion ( usually overload distortion ).

In this chapter, these two problems will be discussed in detail and some solutions will be presented by using both theoretical analysis and computer simulations. Also, the idle channel noise will be discussed in Section 3.6, and in Section 3.7, the design process of low-pass filters and decimators will be described. Finally, some basic simulations will be shown in Section 3.8 and the summary in Section 3.9.

This chapter is not only an attempt of deep understanding and designing of the SDM systems through the theoretical analysis and the computer simulations, but also the base for further research on the adaptive SDM systems which will be described in Chapters 5 and 6.

## 3.2 Structure of the sigma-delta modulator

There are many ways to construct a sigma-delta modulator with a loop filter which depends on a finite set of parameters. What is meant by *structure* of the sigma-delta modulator is the particular way of realising it.

A recent exact analysis of the single-loop ( first order ) sigma-delta modulator [10][11][12] reveals that when the input is either dc or sinusoidal, the spectrum of the binary quantiser noise is discrete and highly coloured. Both the strength and the location of the noise frequency components varies with the input signal level. Candy has introduced a double loop ( second order ) modulator based on the idea of

embedding a sigma-delta loop within the main loop [8]. Subsequently, several researchers extended the idea to even higher order modulators. A general structure for a higher order multi-loop sigma-delta modulator was proposed by Chao et al.[15], which is shown in Fig. 3-1. It has been shown that a higher order ( multi-loop ) sigma-delta modulator increases the signal-to-noise ratio possible for a given oversampling ratio and has a less spiky quantisation noise spectrum. In other words, the noise tends to be more random. However, modulators with more than two loops (higher than second order ) can latch into undesirable noisy modes due to overloading and hence require very careful design and fine tuning.

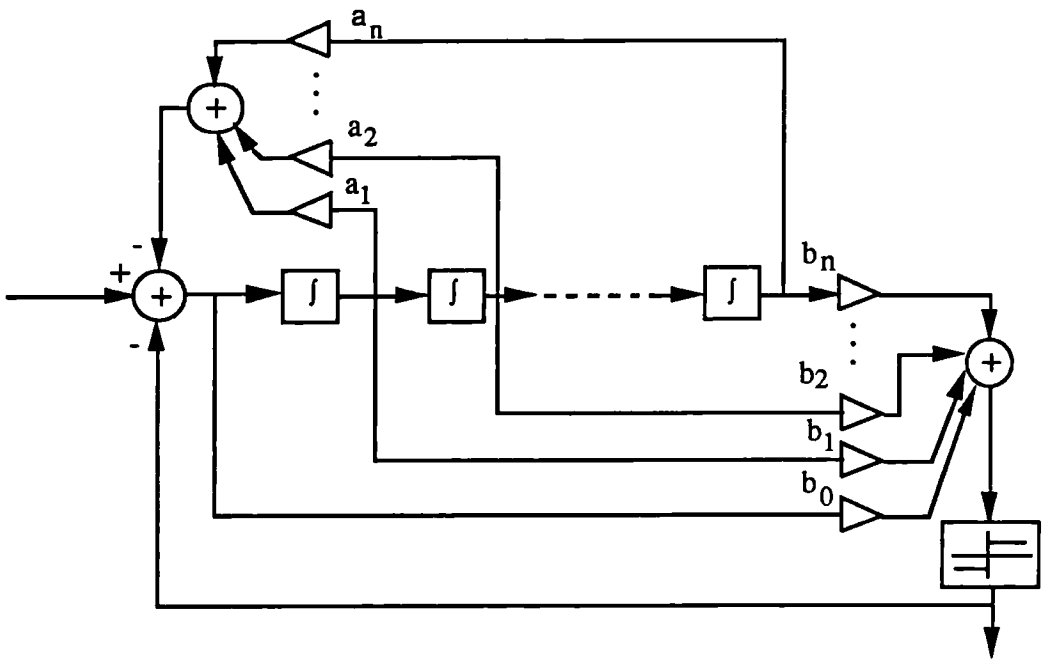



Fig. 3-1 A general structure of a higher order multi-loop sigma-delta modulator proposed by Chao et al.

An alternative structure of a higher order modulator is multi-stage sigma-delta modulation ( also called MASH ), which was proposed by Uchimura et al. [14] [37]. It

consists of cascaded first and/or second order sigma-delta modulators. A structure of a three-stage MASH is shown in Fig. 3-2, in which each stage includes a first order sigma-delta modulator. It has been shown experimentally for the cases of two and three stages that the quantisation noise spectrum of such systems is smooth and that it is free from overloading problems.[4] But the main limitation of MASH structures is their sensitivity to component mismatch between individual stages. It should be noticed that the MASH structure uses more than one quantiser. The number of quantisers used is equal to the number of stages.

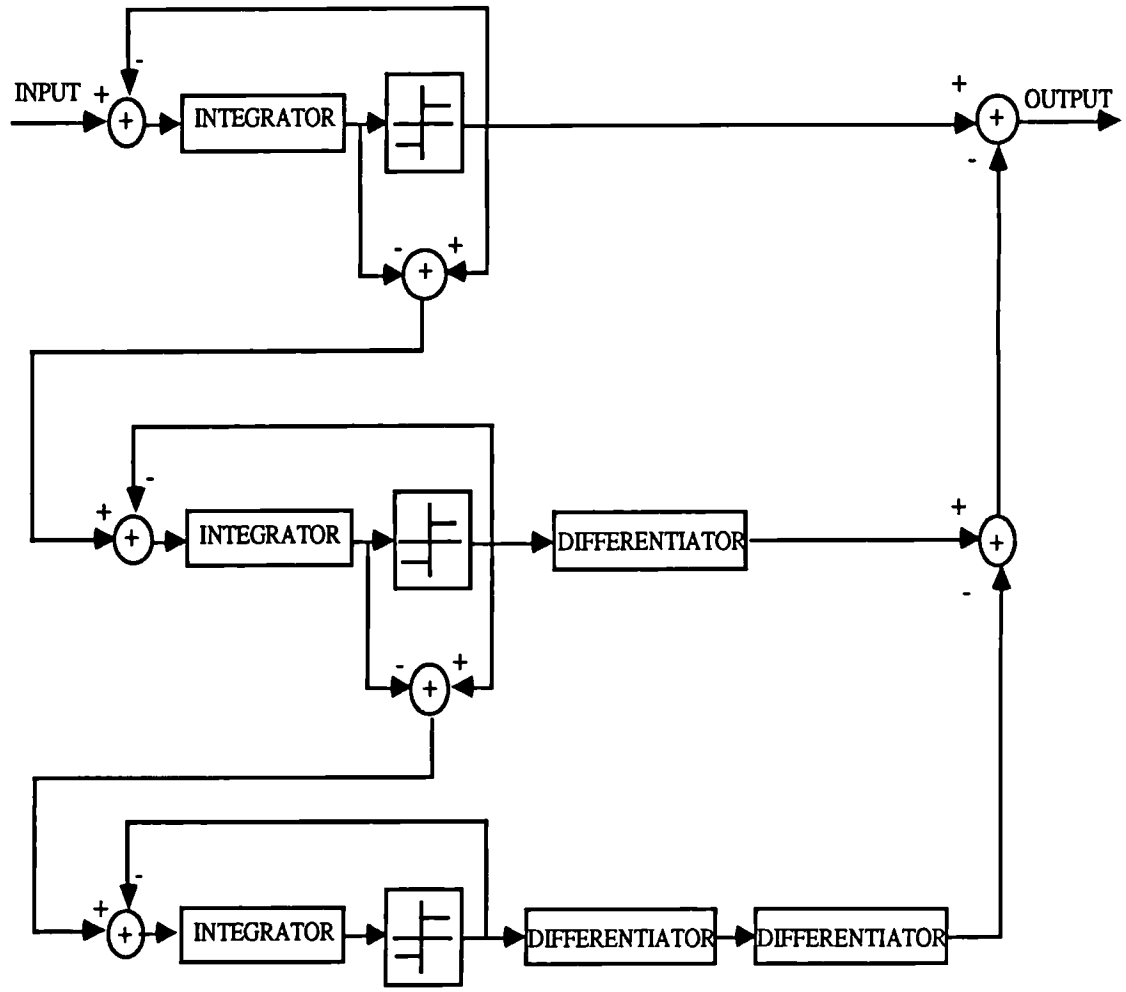



Fig. 3-2 A three-stage MASH structure

The structure of a discrete-time model sigma-delta modulator shown in Fig. 3-3 has been used for all the computer simulations of the project through this thesis, which is mainly based on the structure of Fig. 3-1. Comparing with Fig. 3-1, individual block $z^{-1}/(1-z^{-1})$ plays an equivalent role to an integrator to those in the continuous-time model. It also should be noticed that $b_0$ in Fig. 3-3 is zero. This is because there is at least one delay from the feedback line. The current output cannot directly contribute through the feedback line to the output again without any delay. In [15], a $z^{-1}$ delay is associated with the quantiser for avoiding this problem. From the author's point of view, the structure in Fig. 3-3 seems more sensible.

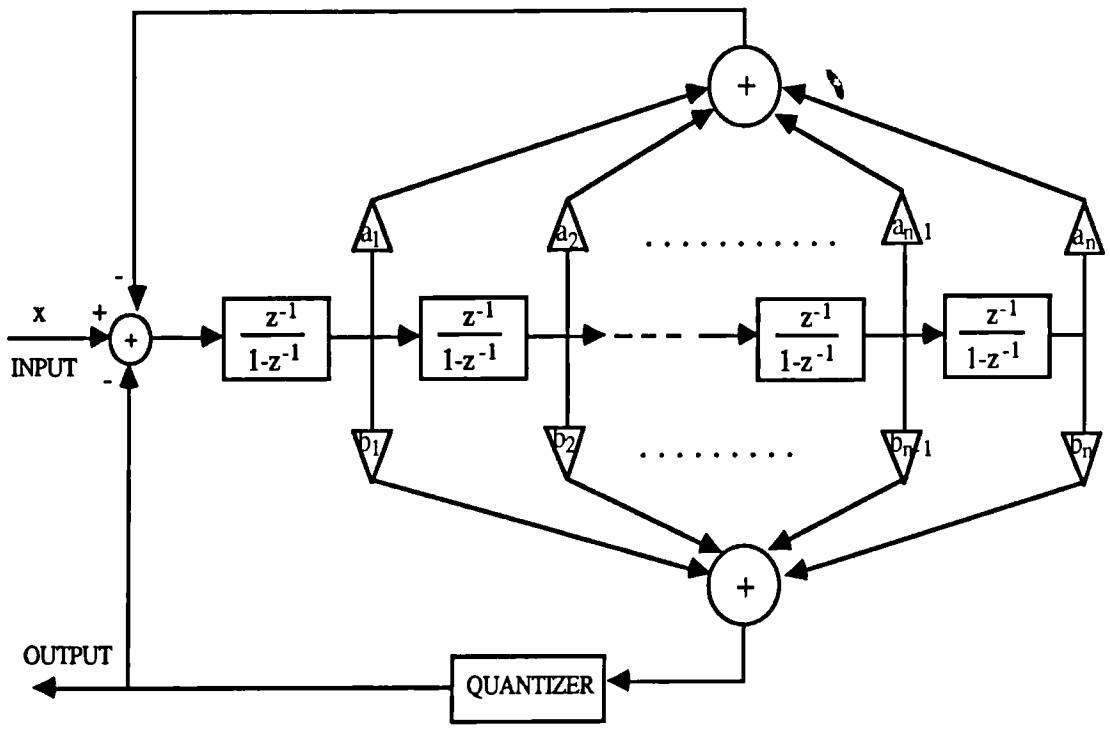



Fig. 3-3 Structure for the nth order SDM

The main reasons for choosing the structure of higher order single stage rather than a MASH structure besides the disadvantage of the MASH mentioned above are:

1) more work needs to be carried out on designing an optimal system of this structure in the sense of stability and SNR;

2) the MASH has more than one quantiser so that it is more difficult to start with in investigating the adaptive quantisation because it will introduce further mismatching problem.

## 3.3 Optimisation of the coefficients of the loop filter

If we represent the loop filter by using G(z) (see Fig.2-1), then G(z) is in the form of

$$G(z) = \frac{b_1(z-1)^{n-1}+b_2(z-1)^{n-2}+...+b_{n-1}(z-1)+b_n}{(z-1)^n+a_1(z-1)^{n-1}+...+a_{n-1}(z-1)+a_n}$$

For a given set of $\{a_i\}$, let $T_{\{b_i\}}[.]$ represents the operation of the system in Fig.3-3 on the input x, the output q is: $q=T_{\{b_i\}}[x]$, then the following property holds.

**Property 3.1** $\forall \{b_i\} \ni \{Kb_i\}$, $K \in (0, \infty)$, there will be $T_{\{b_i\}}[x] = T_{\{Kb_i\}}[x]$

**Proof**

(i) Suppose that an extra gain K is placed after the filter G(z) and before the quantiser, which is equivalent to the multiplication of the $\{b_i\}$ coefficients by K and assume that u is the output of the filter G(z).

(ii) One-bit quantiser has the property $Q(Ku) = Q(u)$, $K \in (0,\infty)$.

(iii) $T_{\{b_i\}}[x] = Q(u)$ and $T_{\{Kb_i\}}[x] = Q(Ku)$

□

The property states that the operation of the whole system will not be affected if all the $b_i$ coefficients are multiplied by a positive real number K. If one set of $\{b_i\}$ is found to be optimal in the sense of stability and maximum signal-to-noise ratio, there will be infinite sets of $\{Kb_i\}$, $K \in (0,\infty)$, which satisfy the same condition of stability and signal-to-noise ratio. But they are all linearly dependent.

There were several papers on designing the filter G(z) [38] [15]. They are mostly based on the concept of linear system so that the signal and noise can be seperated. The noise transfer function $F_E(z)$ is then designed according to the desired shape and the coefficients of the loop filter G(z) can be derived from the relation

$$F_E(z) = \frac{1}{1 + G(z)} \tag{3-1}$$

The concept of the linear system is not applicable to nonlinear systems like single-bit SDM in some aspects. For example, according to the property 3.1, KG(z) plays exactly the same role as G(z). If we use KG(z) instead of G(z), then

$$F_E(z) = \frac{1}{1 + KG(z)}$$

The design of $F_E(z)$ is therefore meaningless. The phenomena of the nonlinearity will be discussed in more detail in the next chapter.

We approach the design from the angle of optimisation of nonlinear systems. The system is considered as a black box which is controlled by the coefficients $\{a_i, b_i, i=1,..n\}$, where n is the order of the system. It is illustrated in Fig. 3-4. The system will introduce some noise in the output. The $\{a_i, b_i\}$ coefficients are optimised in the way that the output approaches the input as "near" as possible. The similarity between the input and the output is measured by signal-to-noise ratio (SNR). The SNR is calculated in the frequency-domain because of the delay problem, which will be explained in Appendix A. Note that the concept of the noise transfer function is not used here. A sinusoidal signal is chosen as the input because it is commonly used and it is easy to measure the SNR of the system by using it. Also, a large amount of signals consist of different sinusoidal frequencies. According to Steele[25], the SNR of the SDM is independent of the signal frequency under the condition of white noise. As a result, the frequency of the sinusoidal input is less important as the order of the loop filter increases. However, some certain frequencies which have integer sub-multiple relationships with the sampling frequency cannot be used [39]. (Unfortunately, dc signal is included. Otherwise, the optimisation programme would have been much simpler.)

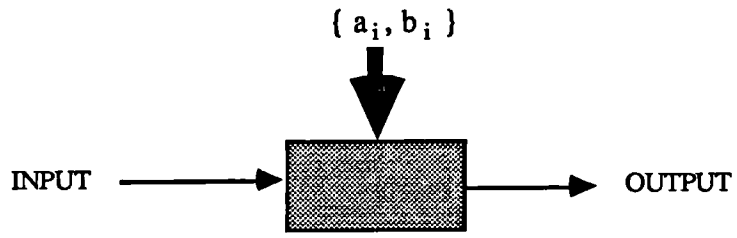



Fig. 3-4 Considering a SDM system as a black box which is controlled by the coefficients $\{a_i, b_i\}$

Let SNR be a function of coefficients $a_i$, $b_i$, defined by

$$SNR = SNR(a_i , b_i)$$

This measurement also includes the stability problem. If the system is unstable under a certain group of $a_i$ and $b_i$, then it will not have a reasonable SNR. The optimisation problem is thus formulated as a maximisation problem. The optimisation function defined by $f_{opt}$ will be

$$f_{opt} = \underset{\substack{\{a_i\},\{b_i\} \\ i=1,\ldots,n}}{\text{maximum}} SNR ( \{a_i\}, \{b_i\} )$$

where n is the order of the system.

Computer simulations and some other researchers[38][15] have shown that the $b_i$ coefficients play the major role in stability and base-band noise suppression whereas the $a_i$ coefficients are not very crucial. Therefore, for designing the filter,to begin with, $a_i$ coefficients are set to be zero and only the optimisation of $b_i$ coefficients is carried out. Furthermore, according to Property 3.1, we can always let $b_1$ equal one and find the optimal set of $\{b_i\}$ under this particular condition and any other set will be easily gained by multiplication by a factor of K. As a result, the problem of n dimensional optimisation of $b_i$ for the nth order SDM will reduce to a (n-1) dimensional one. The optimisation function will be

$$f_{opt} = \underset{\{b_i,\ i=2,\ldots,n\}}{\text{maximum}} SNR ( \{a_i\}=0.0, b_1=1.0, \{b_i\})$$

Once the optimal $\{b_i\}$ have been found, defined as $\{b_{iopt}\}$, the $\{a_i\}$ will be determined

to further improve the signal-to-noise ratio in the base-band. The optimisation function is then changed to

$$f_{opt} = \underset{\{a_i,\ i=1,\ldots,n\}}{\text{maximum}}\ \text{SNR}\ (\ \{b_i=b_{iopt}\},\ \{a_i\}\ )$$

Because the closed form of SNR function is not known, the optimisation methods which require gradient information cannot be used. Although, in some cases, the difference can be used as an approximation of the derivative, in the particular case of SDM system, the results of searching are not satisfying. Another type of method which does not require the partial derivatives of the function can be considered, which are called search methods. They attempt to increase the value of the objective function: SNR by the use of tests near to an estimate of the solution. Two simple methods among them are "simplex" and "pattern" search. The pattern search method is chosen because it can automatically increase or decrease the step size of iteration according to the current result. The detail of the method is described in Appendix B.

It is found that there are many local optimal points which give similar SNR results. However, these points are concentrated in a small particular neighbourhood. Fig. 3-5 shows some of $b_2$-$b_3$ points for the 3rd order SDM.

Table 3-1 gives some of the optimisation results from the 1st order to the 4th order SDMs with oversampling ratio being 64. The floating-point data are used for both the input and the output. Fig. 3-6 shows the comparison between two output spectra of the third order SDM. One is with coefficients $\{a_i\}$ being zero; only $\{b_i\}$ are used. The other one is the result of using both $\{b_i\}$ and $\{a_i\}$. It shows that by only using $\{b_i\}$,

the spectrum is more like conventional one when analysing SDM as a linear system and using $(1-z^{-1})^n$ as a noise transfer function. Unfortunately, the conventional one will cause the system to become unstable when n is greater than two. By using optimisation method, we can reach the same kind of noise curve while maintaining the system stability. The $\{a_i\}$ coefficients reshape the noise curve in a way that the noise is more evenly distributed inside the signal band.

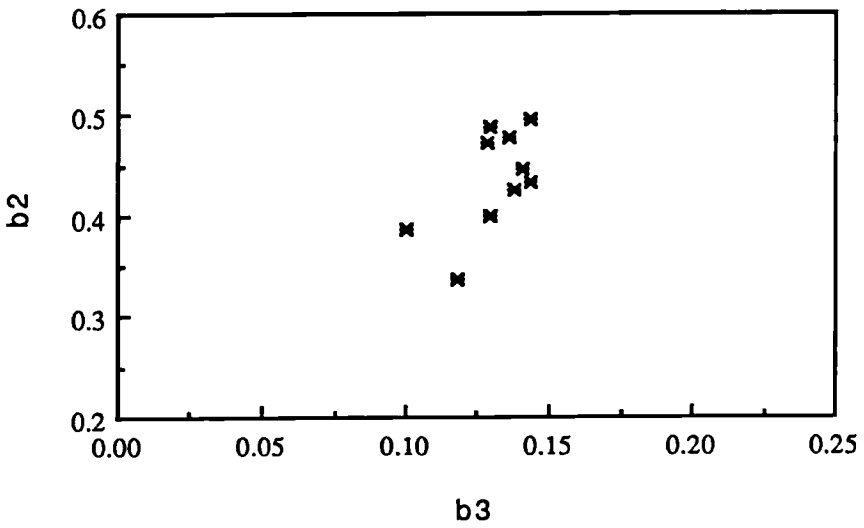



Fig. 3-5 Some optimal $b_2$-$b_3$ points of the 3rd order SDM

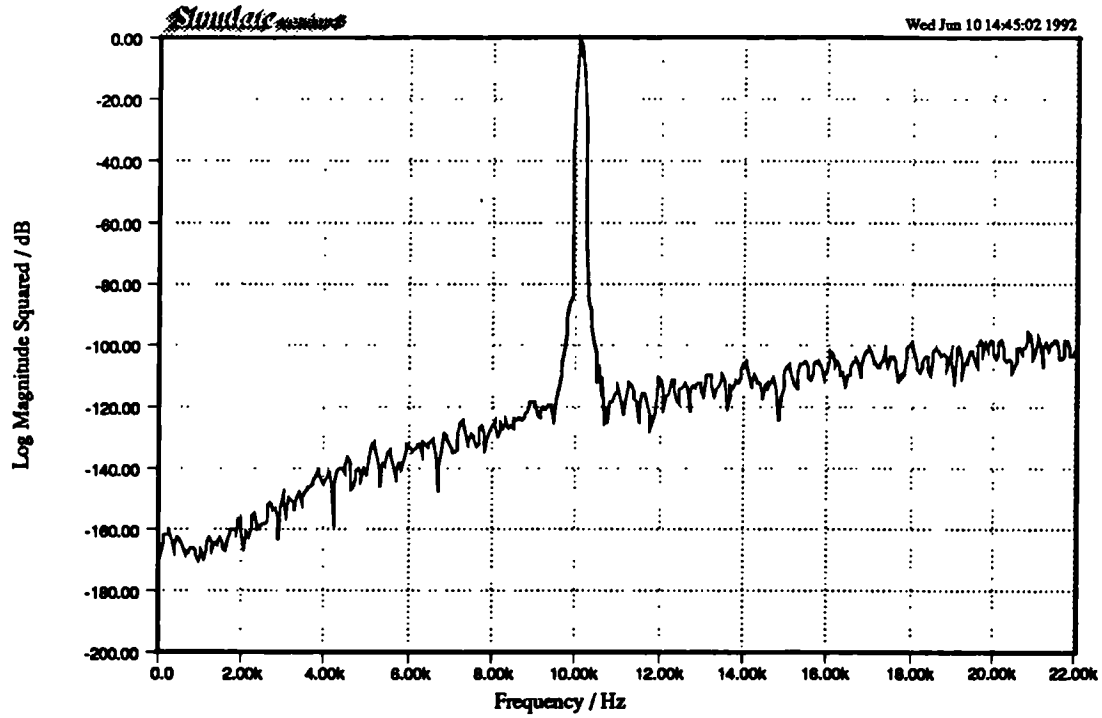
By fixing the optimal $\{a_i\}$ coefficients, $\{b_i\}$ coefficients can be adjusted again by optimisation. Nevertheless, it is found that the SNR can hardly be further improved.

Table 3-1 Simulation results of the filter coefficients

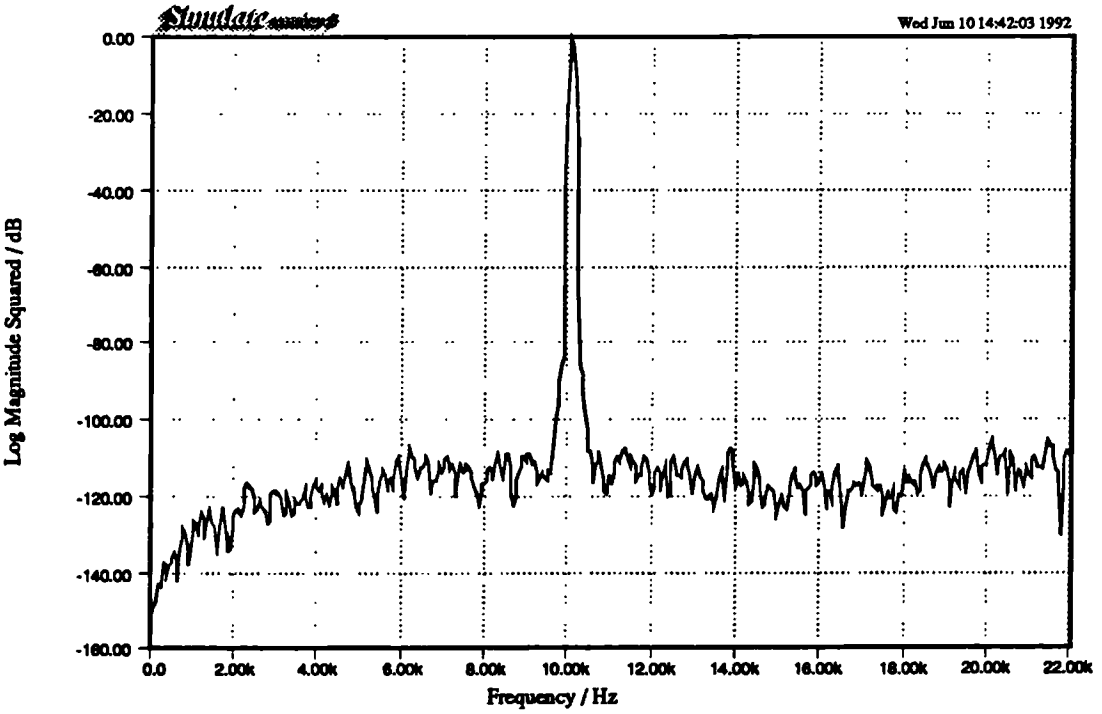
| The Order of the Filter / Coefficients | 1 | 2 | 3 | 3 | 4 |
|---|---|---|---|---|---|
| $b_1$ | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| $b_2$ | — | 0.395625 | 0.5 | 0.447112 | 0.5459 |
| $b_3$ | — | — | 0.1301 | 0.140976 | 0.133846 |
| $b_4$ | — | — | — | — | 0.022432 |
| $a_1$ | 0.0059375 | 0.0028125 | 0.00001 | 0.002 | 0.008564844 |
| $a_2$ | — | 0.00075 | 0.00116 | 0.0012187 | 0.00151 |
| $a_3$ | — | — | 0.0 | 0.0000013458 | 0.000001 |
| $a_4$ | — | — | — | — | 0.0 |
| SNR($a_i = 0$) dB | 52.8 | 71.81 | 84.8 | 85.04 | 93.84 |
| SNR($a_i \neq 0$) dB | 57.2 | 75.9 | 91.0 | 90.86 | 101.94 |

The difference between the signal level and the highest noise level in dB can be seen from Fig. 3-6. However, this difference is not the SNR which is the result of the ratio of signal and noise integrals along the frequency or time axis. The SNR value is worse than this difference. The SNR is about 91 dB for the spectrum of Fig. 3-6(b), but the difference between the maximum signal level and the noise level is more than 100 dB. The SNR measurement will be described in Appendix A.

The *Simulate (version 3)* software package [57] was used to obtain the time and frequency domain plots and to perform FFT based spectral analysis procedures throughout this thesis.




(a) Spectrum of the SDM with $\{a_i\}$ being zero




(b) Spectrum of the SDM with $\{a_i\}$ coefficients

Fig. 3-6 Effect of adding $\{a_i\}$ coefficients on the spectrum
(FFT length: 1024 ; window type: Nuttal [47], see p.140 of chapter 5 of this thesis)

## 3.4 Equivalent quantiser

As is viewed in Section 2.8, a sigma-delta modulator combining with a demodulator can be considered as an equivalent quantiser $Q_x$ to the input of the modulator because the output of the system is actually a kind of quantised version of input x. The idea is demonstrated in a more intuitive way in Fig. 3-7. The concept and the results of this section will be used in the later sections in this chapter and also in Chapter 5 for adaptive sigma-delta modulators.

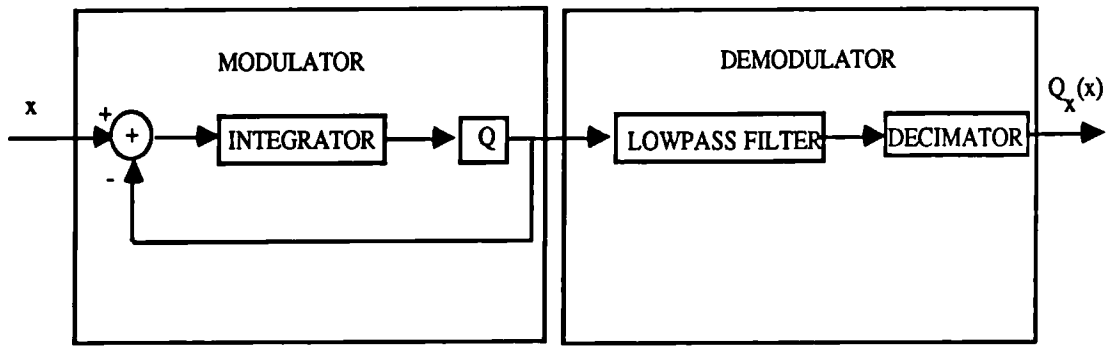



Fig. 3-7 A sigma-delta modulator-demodulator as an equivalent quantiser

Fig. 3-8 shows the block diagram of a discrete time first order sigma-delta modulator and a demodulator with averaging function. The analysis of this simple modulator-demodulator system, when input is dc, has shown that the equivalent quantiser to the input signal x is highly non-uniform. This means that the intervals between the equivalent quantisation levels are different. It can be shown as follows. As is seen in Chapter 2, the output of the equivalent quantiser is

$$Q_x(x) = \frac{1}{N}\sum_{i=0}^{N-1} Q(u_i) = \frac{k_1 - k_2}{N} d \qquad (3\text{-}1)$$

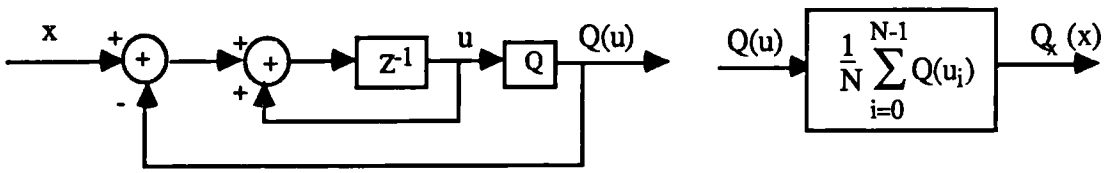
if the averaging function is used as a special case of the demodulator, where $k_1$ and $k_2$ are the number of positive and negative bits among N output bits and d is the quantisation level of the quantiser. From Fig. 3-8, the nonlinear difference equation relating $u_k$ to $u_{k-1}$ is

$$u_k = x - Q(u_{k-1}) + u_{k-1}$$

where x is a constant input. The operation can be summarised by the following difference equations

$$u_k = \begin{cases} u_{k-1} - (d-x), & \text{if } u_{k-1} \geq 0 \\ u_{k-1} + x + d = 2d - (d-x), & \text{if } u_{k-1} < 0 \end{cases} \qquad k = 1,2,\cdots\cdots \qquad (3\text{-}2)$$

where x is assumed to be within the range of [-d, d]. Equation (3-2) states that if $u_{k-1}$ is positive or zero, the state variable is decremented by d-x in the next time step, whereas if $u_{k-1}$ is negative, the state variable is incremented by d+x = 2d - (d-x). Thus,




(a) Modulator (b) Demodulator

Fig. 3-8 A discrete time first order sigma-delta modulator and a demodulator with averaging function

the state variable is always decremented by d-x, but if $u_{k-1}<0$, it is additionally incremented by 2d.

It has been shown by Gray [10] that if the initial integrator state $u_0$ is in the range [x-d, x+d], then the integrator state remains in the same range at all future times, that is

$$u_0 \in [x-d, x+d] \Rightarrow u_k \in [x-d, x+d] \qquad \text{for all } k > 0 \tag{3-3}$$

Furthermore, even if the initial condition u is such that $u_0$ is not within the interval, $u_k$ will progress monotonically toward the interval and eventually lie within it.

Supposing that the numbers of positive and negative bits are $k_1$ and $k_2$ respectively among the first N output bits, to satisfy (3-2) and (3-3), the following condition must be satisfied

$$x - d \le u_0 - N(d-x) + 2k_2d \le x + d \tag{3-4}$$

Considering the simplest case $u_0=0$, then from (3-4) it can be deduced that

$$\frac{k_1-k_2-1}{N-1} d \le x \le \frac{k_1-k_2+1}{N-1} d \tag{3-5}$$

This indicates that when x is within the above range, the N-bit output contains $k_1$ positive and $k_2$ negative bits or we can say that, when the N-bit output contains $k_1$ positive and $k_2$ negative bits, the input x must be within the above range. From (3-1) we also know that when the output sequence contains $k_1$ positive and $k_2$ negative bits, the output of the equivalent quantiser is $(k_1-k_2)d/N$. By calculating the differences between the equivalent quantisation point $(k_1-k_2)d/N$ and the upper bound and the lower bound of (3-5), it is found that the two distances can be different and depend on

the number of positive or negative bits. We shall define the distance between the quantisation point and upper bound as $D_{up}$ and for lower bound, $D_{low}$. As a result,

$$D_{up}=\frac{k_1-k_2+1}{N-1}d-\frac{k_1-k_2}{N}d=\frac{2k_1d}{N(N-1)}$$

$$D_{low}=\frac{k_1-k_2}{N}d-\frac{k_1-k_2-1}{N-1}d=\frac{2k_2d}{N(N-1)} \qquad (3\text{-}6)$$

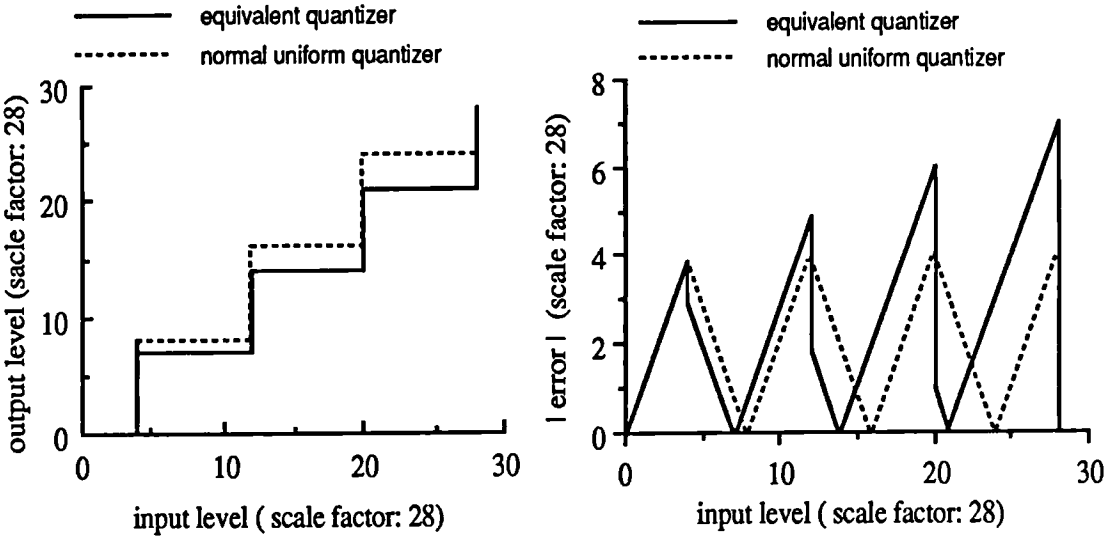
Although the interval widths of the input are the same, and the widths between two adjacent quantisation levels are the same, in general, the output points are not the midpoints of the set of inputs which yield these points as they are in the usual uniform quantiser. They depend on the magnitude of the input. Table 3-2 gives the characteristic of the equivalent quantiser when N=8, which is compared with the normal midtread uniform quantiser whose quantisation level is represented by $Q_u(x)$. Note that when N is even, the value of $k_1-k_2$ is also even. Fig. 3-9 gives the corresponding curves and the curve of quantisation error versus input level comparing with the case of uniform quantiser. It can been seen that the characteristics looks like the same kind as that of uniform quantiser. However, from the error curves, differences between them become obvious. The error curve for the normal uniform quantiser is continuous, but there are some discontinuous points for the equivalent quantiser. For example, when the input level is slightly less than 20/28, the error is 6/28, but when the input level is slightly greater than 20/28, the error will drop to 1/28. The error will reach the maximum value 7/28 when the input is very near to the full range. When it reaches the full scale, the error will drop to zero.

If $u_0 \neq 0$, the curves in Fig. 3-9(a) will shift left or right depending on the value of $u_0$. The interesting thing is that for each block of N samples, the value of the first sample $u_0$ can be different. This indicates that the curve of quantisation level against input level will change from block to block. In other words, the equivalent quantiser is a time varying quantiser.

It should be noticed that all the results above only depend on N and the difference between $k_1$ and $k_2$ no matter how $k_1$ positive and $k_2$ negative bits are distributed among N bits. It should also be noticed that there are only N+1 quantisation levels for N-bit output, but for N-bit PCM, there are $2^N$ levels.

Table 3-2 Characteristic of the equivalent quantiser compared with the normal midtread uniform quantiser

| $k_1 - k_2$ | x | $Q_x(x)$ | $Q_u(x)$ |
|---|---|---|---|
| 0 | (-1/7, 1/7) | 0 | 0 |
| 2 | (1/7, 3/7) | 2/8 | 2/7 |
| 4 | (3/7, 5/7) | 4/8 | 4/7 |
| 6 | (5/7, 7/7) | 6/8 | 6/7 |
| 8 | 7/7 | 8/8 | 6/7 |




(a) Characteristic of the equivalent quantiser (b) Absolute quantisation error

Fig. 3-9 Comparison between the equivalent and normal uniform quantisers

## 3.5 Maximum possible input level and optimal quantisation level for sinusoidal inputs

In the previous section, the results are based on the condition of dc input. In this section, we analyse the system when input is sinewave signal. Supposing the input is a sinusoidal signal with amplitude A: $x_k$=Asin$t_k$ and that N is large enough so that the changes are small over N samples and the following still holds.

$$\frac{k_1-k_2-1}{N-1}d \le x \le \frac{k_1-k_2+1}{N-1}d \quad \Rightarrow \quad Q_x(x) = \frac{k_1-k_2}{N}d \tag{3-7}$$

It is possible now to calculate the signal-to-noise ratio of the simple system in Fig. 3-8. The signal power $x_p$ and the noise power $n_p$ are

$$x_p = \frac{1}{N}\sum_{k=0}^{N-1} A^2 \sin^2 t_k \quad , \qquad n_p = \frac{1}{N}\sum_{k=0}^{N-1} [\, A \sin t_k - Q_x(x_k)\,]^2$$

For the convenience of calculation, we use continuous time function and integration instead of discrete time function and summation. When N is large enough, this approximation is acceptable. Therefore,

$$x_p = \frac{1}{2\pi}\int_0^{2\pi} A^2 \sin^2 t \, dt = \frac{A^2}{2}$$

Because $Q_x(x)$ is a piecewise linear function of x and x is a function of t, the integral along the t axis can be divided into different time intervals. Suppose that N is even (this is the usual case) so that the possible value for $k_1$-$k_2$ is 0,±2,±4..... ±(N-2). Let $i=k_1-k_2-1$ and d=1 (normalised case) in (3-7) and only $x \geq 0$ is considered, when x is within the range

$$\frac{i}{N-1} \leq x \leq \frac{i+2}{N-1}$$

t will be

$$\sin^{-1}\left(\frac{i}{N-1}\right) \leq t \leq \sin^{-1}\left(\frac{i+2}{N-1}\right)$$

and $Q_x(x) = (i+1)/N$. The noise power should be

$$n_p = \frac{1}{2\pi}\left\{4\int_0^{\frac{\pi}{2}} [\, A\sin t - Q_x(A\sin t)\,]^2 \, dt\right\}$$

$$= \frac{2}{\pi}\int_0^{\frac{\pi}{2}} A^2\sin^2 t - 2AQ_x(A\sin t)\sin t + [\, Q_x(A\sin t)\,]^2 \, dt$$

$$= \frac{2}{\pi}\left\{\frac{A^2\pi}{4} - 2A\sum_{i=1,3,\ldots N-3}\frac{i+1}{N}\int_{\sin^{-1}(\frac{i}{N-1})}^{\sin^{-1}(\frac{i+2}{N-1})}\sin t\,dt + \sum_{i=1,3,\ldots N-3}\left(\frac{i+1}{N}\right)^2\int_{\sin^{-1}(\frac{i}{N-1})}^{\sin^{-1}(\frac{i+2}{N-1})}1\,dt\right\}$$

$$= \frac{2}{\pi}\left\{\frac{A^2\pi}{4} + 2A\sum_{i=1,3,\ldots N-3}\left[\cos\left(\sin^{-1}\frac{i+2}{N-1}\right) - \cos\left(\sin^{-1}\frac{i}{N-1}\right)\right]\frac{i+1}{N} + \sum_{i=1,3,\ldots N-3}\left[\sin^{-1}\frac{i+2}{N-1} - \sin^{-1}\frac{i}{N-1}\right]\left(\frac{i+1}{N}\right)^2\right\}$$

Let

$$a_1 = 8\sum_{i=1,3,\ldots N-3}\left[\cos\left(\sin^{-1}\frac{i+2}{N-1}\right) - \cos\left(\sin^{-1}\frac{i}{N-1}\right)\right]\frac{i+1}{N}$$

$$a_0 = 4\sum_{i=1,3,\ldots N-3}\left[\sin^{-1}\frac{i+2}{N-1} - \sin^{-1}\frac{i}{N-1}\right]\left(\frac{i+1}{N}\right)^2$$

then

$$n_p = [\,A^2 + a_1A + a_0\,]/2$$

The signal-to-noise ratio is therefore

$$SNR = \frac{x_p}{n_p} = \frac{A^2}{A^2+a_1A+a_0} = \frac{1}{1+\frac{a_1}{A}+\frac{a_0}{A^2}} \qquad (3\text{-}8)$$

$$\text{Let } C = \frac{1}{A} \quad \Rightarrow \quad SNR = \frac{1}{1+a_1C+a_0C^2} \qquad (3\text{-}9)$$

From (3-8) or (3-9) it can be determined at what value of input amplitude A, the signal-to-noise ratio reaches maximum. In order to make SNR maximum, $1+a_1C+a_0C^2$ should be minimum. It is easy to derive from letting the derivative be zero so that

$$C=C_{opt}= -a_1/(2a_0) \quad \Rightarrow \quad A=A_{opt}= -2a_0/a_1$$

where $C_{opt}$ and $A_{opt}$ are the optimal values of C and A for maximum SNR. It can be shown that $a_0$ ( the second derivative ) is always greater than zero so that $1+a_1C+a_0C^2$ has minimum value and therefore SNR has maximum value. Table 3-3 shows the results of $A_{opt}$, $C_{opt}$ and SNR for different oversampling ratio N.

As we know for N bit linear PCM system, the maximum signal-to-noise ratio occurs when the amplitude A is full scale that is, equal to one in our case. However, from Table 3-3 it shows that $A_{opt}$ is always less than one, although as N increases, $A_{opt}$ approaches one. If as is the case in the usual coding system like PCM, we consider the situation in which A is greater than $A_{opt}$ as overload, then, for a SDM system, the possible maximum amplitude without causing overload is $A_{opt}$ and always less than one. Fig. 3-10 shows the corresponding curve.

Table 3-3 Results of $A_{opt}$, $C_{opt}$ and SNR for different oversampling ratio N (quantisation level d=1)

| Oversampling Ratio N | $C_{opt}$ | $A_{opt}$ | SNR (dB) (A=1) | $SNR_{max}$ (dB) ($A=A_{opt}$) |
|---|---|---|---|---|
| 128 | 1.0086 | 0.9915 | 39.39 | 43.69 |
| 96 | 1.0116 | 0.9885 | 36.81 | 41.15 |
| 64 | 1.0178 | 0.9825 | 33.16 | 37.57 |
| 48 | 1.0243 | 0.9763 | 30.56 | 35.01 |
| 32 | 1.0379 | 0.9635 | 26.86 | 31.37 |
| 24 | 1.0522 | 0.9504 | 24.22 | 28.77 |
| 16 | 1.0830 | 0.9234 | 20.46 | 25.04 |
| 12 | 1.1168 | 0.8954 | 17.77 | 22.34 |
| 8 | 1.1940 | 0.8375 | 13.93 | 18.40 |

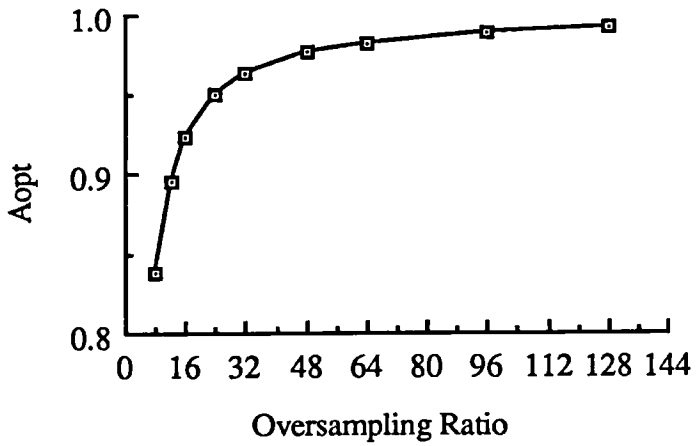



Fig. 3-10 $A_{opt}$ curve versus oversampling ratio

When using a system which is different from Fig. 3-8, the values in Table 3-2 will be different. When the structure of Fig. 3-3 is used, the computer simulation results have shown that the higher the system order, the smaller the $A_{opt}$ and the larger the $C_{opt}$. When the decimator in Fig. 2-11 is used instead of that in Fig. 3-8, the maximum SNR is improved dramatically. For example, when oversampling ratio N is 64, in the case of the 1st order SDM, the maximum SNR increases from 37.57 dB in Table 3-2 to about 53 dB.

If we fix the maximum input amplitude as one for all kinds of SDM systems, then the optimal quantisation levels by which the maximum SNR can be gained for different oversampling ratio and different order of the loop filter will be different and should be equal to $C_{opt}$. Fig. 3-11 shows the results of C values of different order of the system with oversampling ratio being 64. The structures of modulator in Fig. 3-3, the coefficients in Table 3-1, and demodulator in Fig. 2-11 have been used for this

simulation. The results show that the values of $C_{opt}$ are greater than one,

$$C_{opt} = \begin{cases} 1.15, & \text{1st order SDM} \\ 1.5\ , & \text{2nd order SDM} \\ 3.5, & \text{3rd order SDM} \end{cases}$$

which is consistent with the above theoretical analysis. The corresponding values of $A_{opt}$ will be

$$A_{opt} = \begin{cases} 0.87, & \text{1st order SDM} \\ 0.67, & \text{2nd order SDM} \\ 0.29, & \text{3rd order SDM} \end{cases}$$

Also, the value of $C_{opt}$ increases with the order of the system.

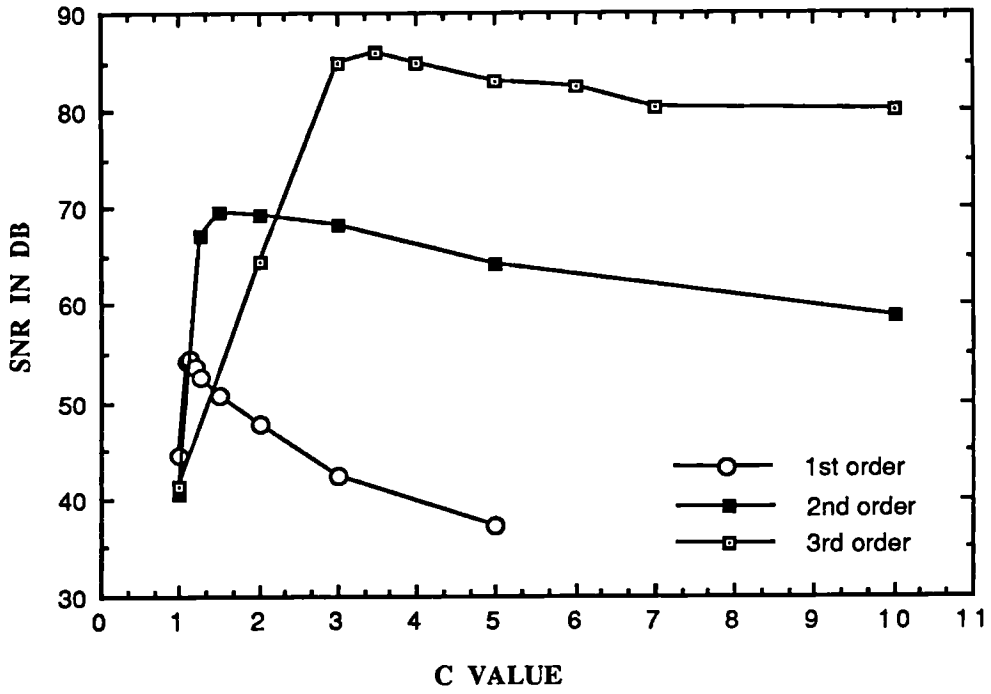



Fig. 3-11 Simulation results of C values for different order of system with oversampling ratio being 64

If we choose an arbitrary value for d as the quantisation level rather than the normalised value 1, then the maximum input magnitude will be $dA_{opt}$. Or we may say

that, if the required maximum input magnitude is $A_{max}$, then the quantisation level should be $d=C_{opt}A_{max}$.

## 3.6 Idle channel noise

Idle channel noise is the coding noise with a zero input. From Fig. 3-9 which is depicted for $u_0=0$, it can be seen that when the input is zero the output will be zero as well. It indicates that the idle channel noise equals zero. Unfortunately, in some situations, even when the input is zero, the output will not be zero. For example, if oversampling ratio N is odd, then $k_1$ can never be equal to $k_2$ in equation (3-1), and hence, $Q_x(x)$ has no zero-level. The smallest level will be d/N. Therefore, in this case, the idle channel noise power will be $d^2/N^2$. In practice, even when N is even, the idle channel noise is never exactly zero, due to some background noise. In addition, practical quantiser characteristics can be non-ideal and asymmetrical. As is mentioned in Section 3.4, the values of $u_0$ for each block of N samples can be different. Therefore, the curve in Fig. 3-9(a) will shift left or right depending on $u_0$. The maximum shift can be 2d/N if $u_0 \in [-d,d]$. The idle channel property can be generally analysed as follows.

If input x=0, for the simple 1st order system in Fig. 3-8, equation (3-3) can be described as

$$u_0 \in [-d, d] \Rightarrow u_k \in [-d, d], \quad \text{for all } k > 0$$

Equation (3-4) can be expressed as

$$-d \le u_0 - Nd + 2k_2 d \le d$$

Consider $N=k_1+k_2$, so that

$$u_0 + d \ge ( k_1 - k_2 ) d \ge u_0 - d$$

Supposing $u_0 \in [-d, d]$, the following will be true

$$| ( k_1 - k_2 ) d | \le 2d \qquad (3\text{-}10)$$

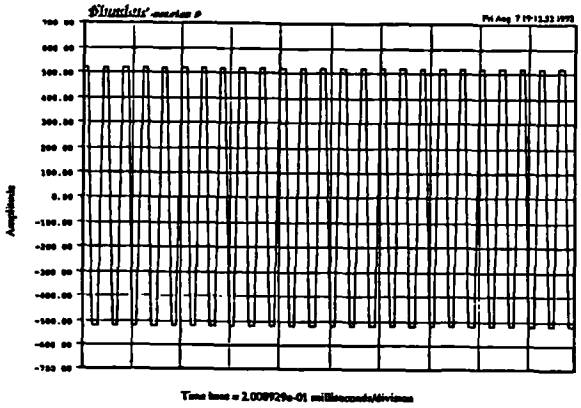
From (3-1) we know that the output of the equivalent quantiser is $Q_x(x)=(k_1-k_2)d/N$. In the case of non-zero idle channel noise, $k_1 \neq k_2$. From (3-10), the idle channel noise will satisfy

$$| Q_x(x) | = | \frac{k_1-k_2}{N} d | \le \frac{2}{N} d \qquad (3\text{-}11)$$

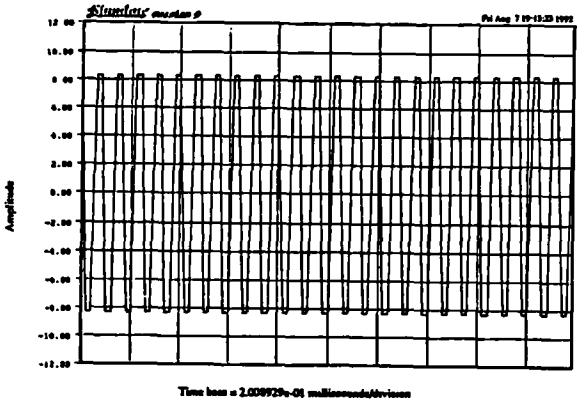
The upper bound of the idle channel noise is proportional to the quantisation level and inversely proportional to the oversampling ratio. The larger oversampling ratio can lead to smaller idle channel noise. The smaller the quantisation level d, the smaller the noise. However, d must be large enough to prevent the overload distortion. This conflict cannot be solved by a fixed quantiser. In Chapter 5, we will discuss how an adaptive quantiser can be used to reduce the idle channel noise as well as avoiding the overload distortion.

The above analysis is based on using one stage of comb filter. If the number of the stages, n, in equation (2-34) increases, the idle channel noise will decrease. Fig.3-12 shows the waveforms of the idle channel noise of a second order SDM with three different demodulators: one, two, and three stage comb filters, where the value of quantisation level d is 32767. The idle channel noise can be reduced by about 36 dB by each stage of comb filter.

(a)



(b)



(c)

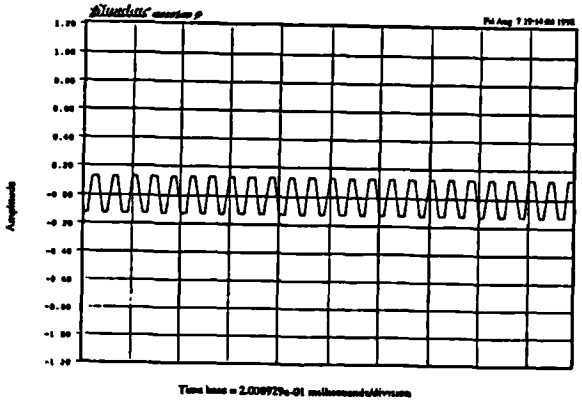


Fig. 3-12 Waveforms of the idle channel noise: (a) using one stage of comb filter; (b) using two stage of comb filters; (c) using three stage of comb filters

## 3.7 Design of decimators

As is mentioned in Chapter 2, the simplest and most economical filter to reduce the input sampling rate is a comb filter, because such a filter does not require a multiplier. This comb filter operation is equivalent to a rectangular window finite impulse response (FIR) filter. However, the comb filter is not very effective at removing the large volume of out-of-band quantisation noise generated by the sigma-delta modulators. Also, the frequency response of the comb filter can cause substantial magnitude drooping at the upper region of the baseband. Therefore, it is seldom used in practice without additional digital filters.

The structure in Fig. 2-11 is chosen for the decimator. A (n+1)-stage comb filter is used to decimate the output of the nth order sigma-delta modulator. The second and third sections are FIR low-pass filters with symmetric coefficients to maintain a linear-phase response. A cascaded half-band filter structure has the following advantages: significantly reduced number of computations; reduced memory requirement; simplified filter design problem; and reduced finite-word-length effects [40].

The method for designing symmetric, half-band FIR filters has been used for filter F1(z) and F2(z). If we consider the special case

$$\delta_s = \delta_p = \delta$$

$$\omega_s = \pi - \omega_p$$

where

$\delta_s$ - the stop band ripple $\delta_p$- the pass band ripple

$\omega_p$- the pass band edge frequency $\omega_s$- the stop band edge frequency

then the resulting equiripple optimal solution has the property that [40]

$$H(e^{j\omega}) = 1 - H(e^{j(\pi-\omega)}) \tag{3-12}$$

That is, the frequency response of the optimal filter is symmetric around $\omega=\pi/2$, it can be derived,

$$H(e^{j\frac{\pi}{2}}) = 0.5$$

It can also be readily shown that any symmetric FIR filter satisfying (3-12) also satisfies the ideal time domain constraints[40]

$$h(k) = \begin{cases} 1, & k=0 \\ 0, & k=\pm 2, \pm 4, \ldots\ldots \end{cases} \tag{3-13}$$

That is, every other impulse response coefficient (except for k=0) is exactly zero. Thus, a factor of two reduction in computation is obtained.

It can be a much wider transition band for the first half-band filter F1(z) than for the second F2(z). Suppose that the sampling rate for F1(z) is 4 Hz and for F2(z) is 2 Hz. Then, for the first half-band filter F1(z), the design specifications can be

$$\delta_{s1} = \delta_{p1} = 0.00001, \quad f_{p1} = 0.5, \; f_{s1} = 1.5$$

and for the second

$$\delta_{s2} = \delta_{p2} = 0.000001, \quad f_{p1} = 0.455, \; f_{s1} = 0.545$$

The transition band specifications are illustrated in Fig.3-13.

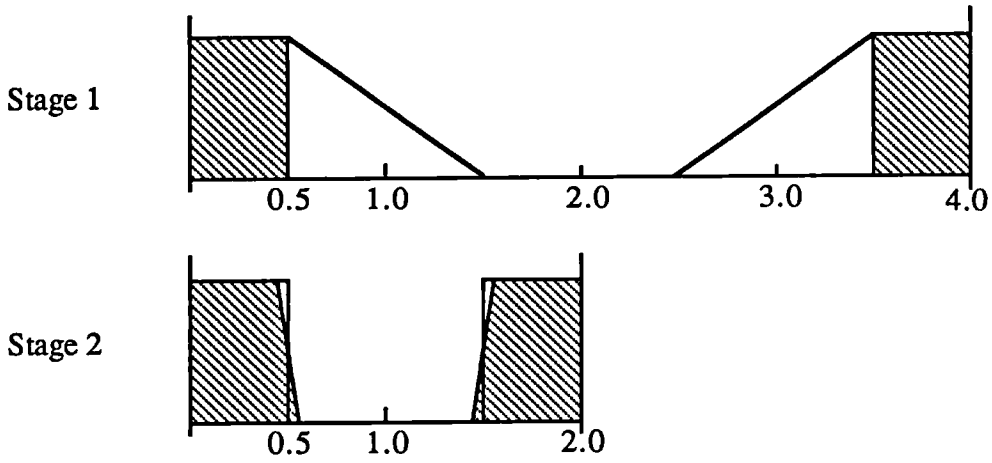



Fig. 3-13 Decimation process of a two-stage half-band filter

By using a standard design programme,[58] the obtained lengths of F1(z) and F2(z) are 25 and 169 respectively. After converting the final results into the form which satisfies time domain constrains (3-13), the final frequency response are shown in Fig.3-14 and 3-15, where the pass-band ripples are

$$\delta_{p1} = 0.0000058 \qquad \delta_{p2} = 0.00000009$$

and the stop-band attenuations are

$$\delta_{s1} = -104 \text{ (dB)} \qquad \delta_{s2} = -115 \text{ (dB)}$$

The combination of the two filters and four stages of comb filter gives the resolution about 118 dB of SNR. This is tested by passing the very high resolution sinewave through the filters while oversampling ratio is 64. Therefore, the decimator can be used for as high as 16-18 bit PCM quality.

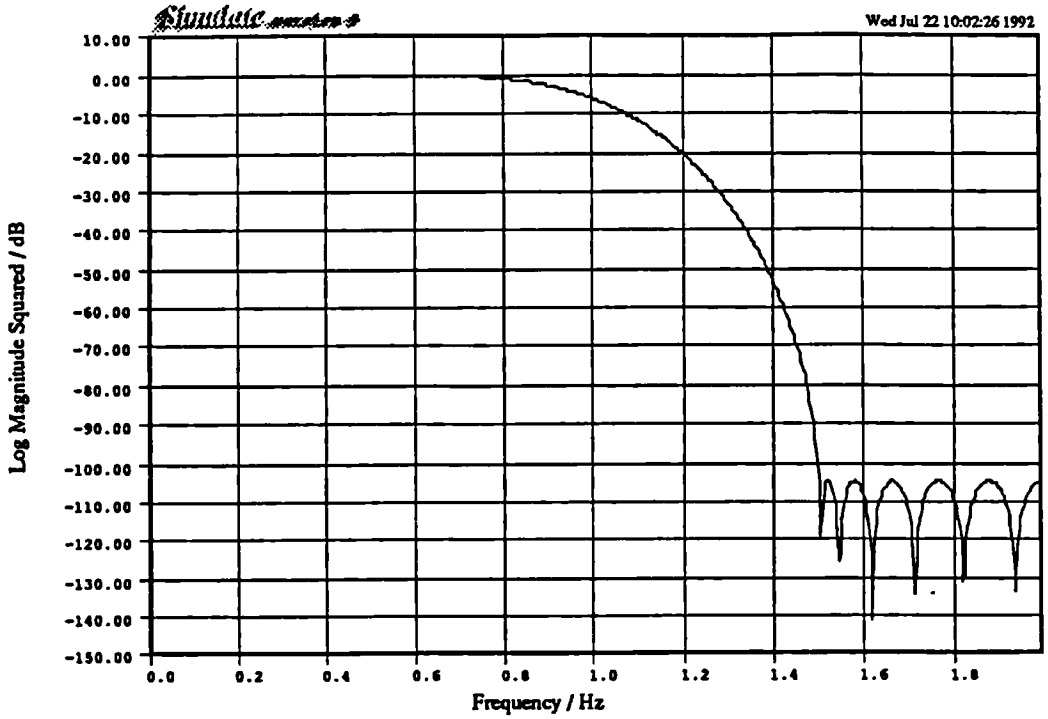



Fig. 3-14 Magnitude response of the 25th order half-band low-pass filter

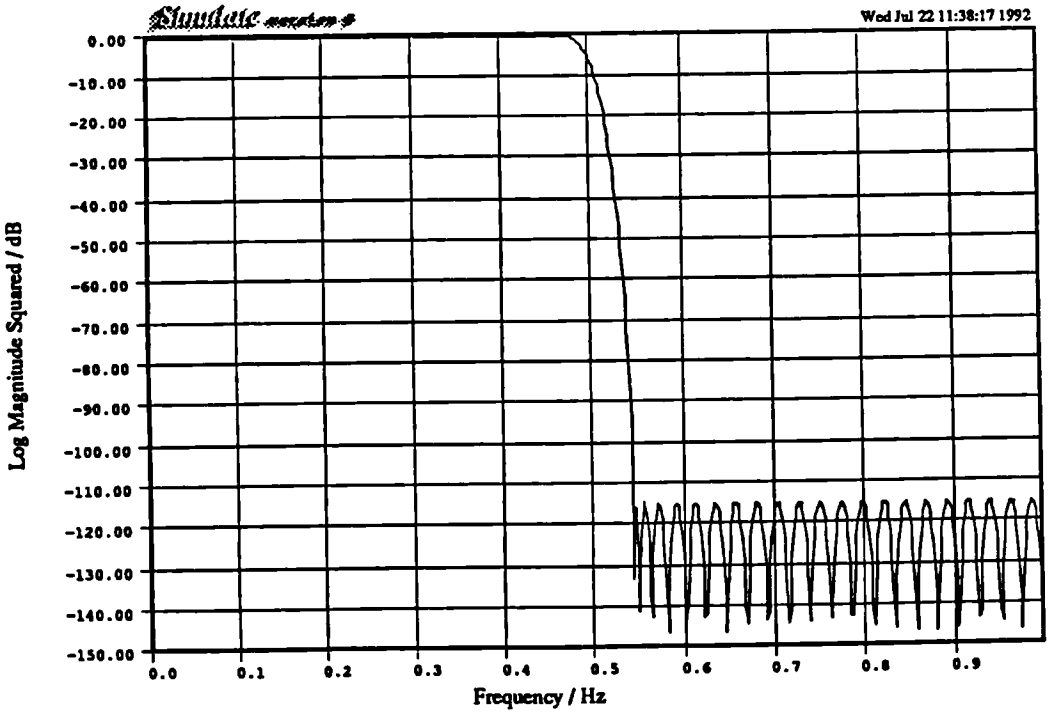



Fig. 3-15 Magnitude response of the 169th order half-band low-pass filter

## 3.8 SNR simulations for sigma-delta modulators

From equation (2-22), the approximate theoretical results of the maximum signal-to-noise ratio for different bit number, oversampling ratio, and order of loop filter can be obtained. The equations were derived based on the assumption of an independent white noise model. The computer simulations show that the equations (2-22) match better with the simulations of lower order and higher bit number systems. Fig. 3-16 gives the curves of maximum signal-to-noise ratio versus bit number when the oversampling ratio is 64. The comparison between the theoretical and the simulation results shows that the difference becomes smaller for multi-bit systems. The reason is probably that the model of independent white noise is more accurate for the system with more quantisation levels. Fig. 3-17 shows the difference between the theoretical values and the simulation results when varying the order of the loop filter. This difference becomes larger when the order increases. This is probably because the higher order loop filter increases the magnitude of the signal before the quantiser so as to cause a much more severe nonlinearity.

Fig. 3-18 gives the results of maximum signal-to-noise ratio of one-bit second order sigma-delta modulation with different oversampling ratios. The difference between the theoretical and simulation results is almost constant for different oversampling ratio. It indicates that the signal-to-noise ratio can be approximately obtained by using equations (2-22) and then subtracting the difference.

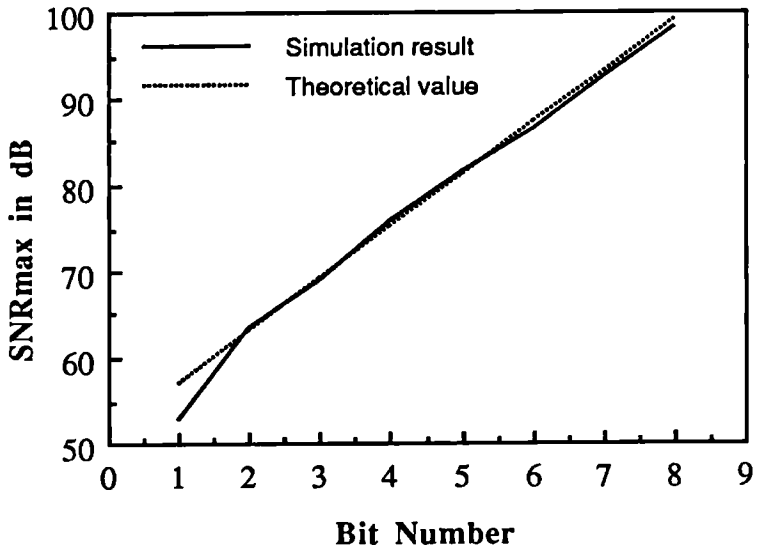



Fig. 3-16 Maximum signal-to-noise ratio versus bit number

( oversampling ratio: 64 frequency: 10087 Hz )

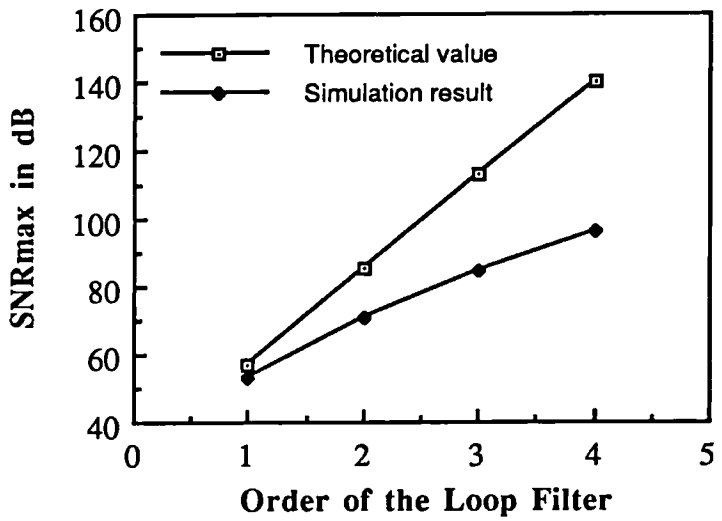



Fig. 3-17 Maximum signal-to-noise ratio versus order of the loop filter

( oversampling ratio: 64, frequency: 10087 Hz , one-bit quantiser)

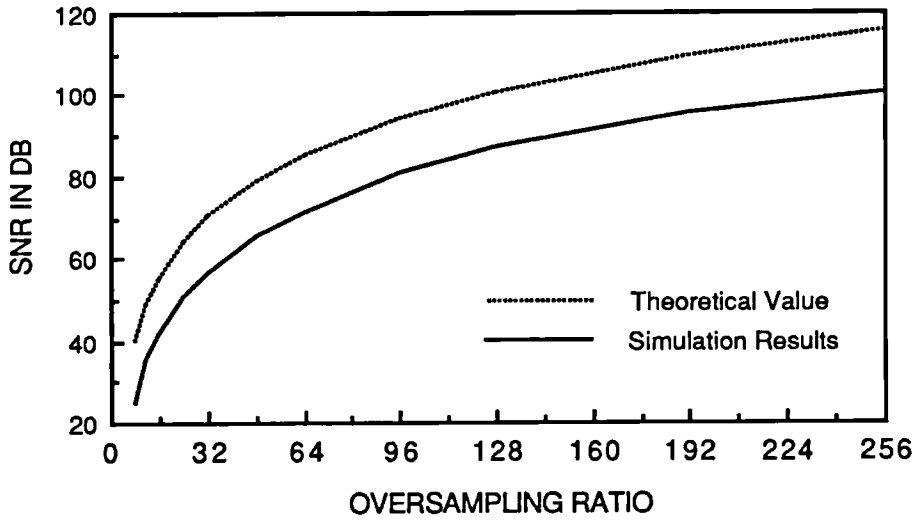



Fig. 3-18 Maximum signa-to-noise ratio of one-bit second order SDM versus oversampling ratio ( frequency: 10087 Hz )

Fig. 3-19 gives the curves of maximum SNR versus oversampling ratio for the 1st, 2nd, and 3rd order one-bit SDM. As is expected, the higher the order and the higher the oversampling ratio, the better SNR the system has. It can also be seen that for a certain oversampling ratio, the difference between the 1st and 2nd, the 2nd and 3rd order systems in SNR are almost the same. By calculating the differences from (2-22), it can be determined that

$$diff_{2,1}\ (dB) = 6.02L - 8.0$$

$$diff_{3,2}\ (dB) = 6.02L - 8.48$$

$$diff_{4,3}\ (dB) = 6.02L - 8.85$$

where $2^L$ is the oversampling ratio and $diff_{i,j}$ is the difference between the ith order and the jth order SDM in SNR. When L is constant, there is less than 0.5 dB difference.

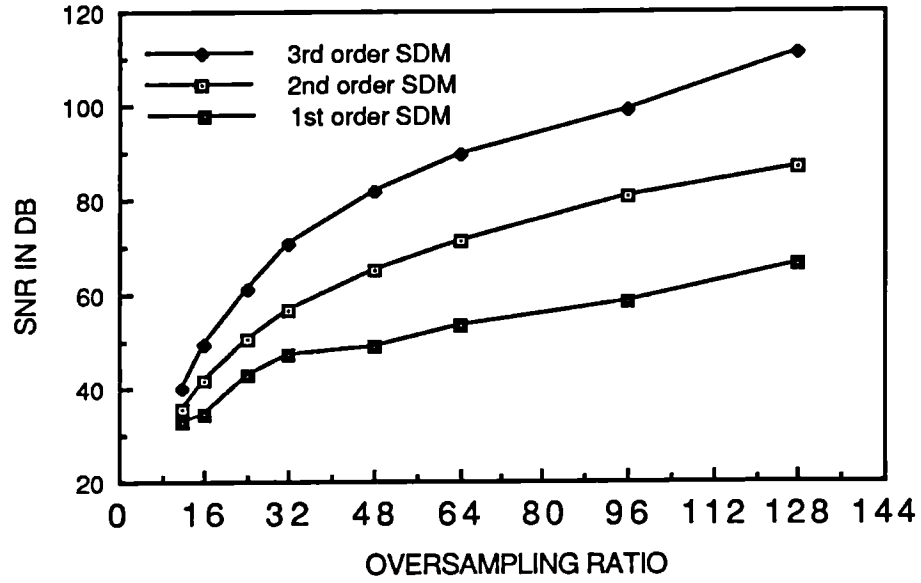



Fig. 3-19 Maximum signal-to-noise ratio of the 1st, 2nd, and 3rd order one-bit SDM versus oversampling ratio ( frequency: 10087 Hz )

Fig. 3-20 gives the SNR curve versus input magnitude of the 3rd order SDM. The level 0 dB corresponds to the maximum possible magnitude of the input. This curve is very similar to the curve in Fig. 2-2, which shows that a SDM used as a A/D converter has the same characteristic as a linear (uniform) A/D converter has.

As is mentioned in Chapter 2, one of the superior properties of SDM over DM is that the overload characteristic is independent of the frequency of the input signal. This has been tested by varying the frequency of the sinusoidal input and measuring the maximum SNR. The result is shown in Fig. 3-21.

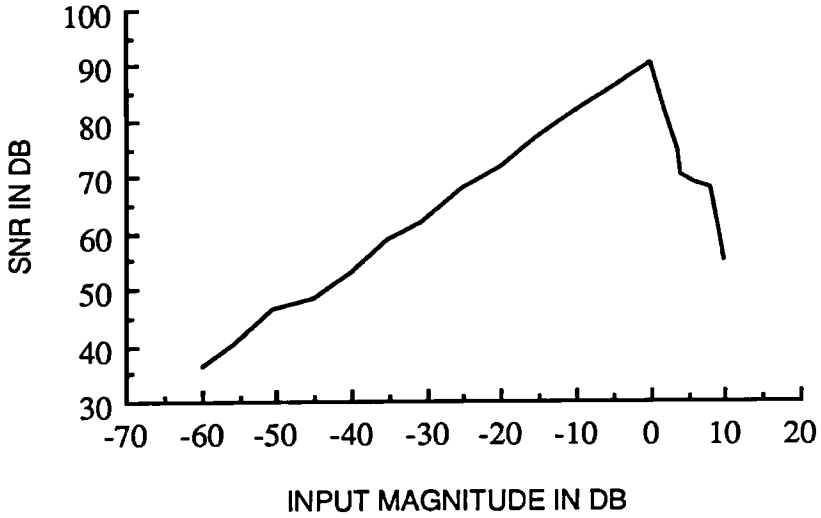



Fig. 3-20 SNR curve versus input magnitude of the 3rd order SDM

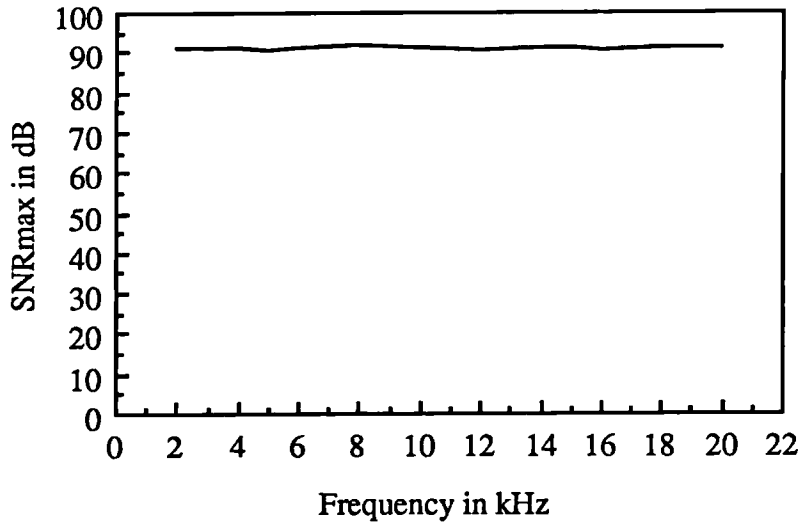



Fig. 3-21 Maximum signal-to-noise ratio versus the frequency of the sinusoidal input

## 3.9 Summary

The key points from this chapter are summarised as follows

1. The coefficients $\{b_i\}$ of the loop filter affect both the stability and the SNR of the SDM system. The coefficients $\{a_i\}$ of the loop filter affect the SNR but have little effect on the stability.

2. If an extra gain K is placed between the filter G(z) and the quantiser, which is equivalent to the multiplication of the $\{b_i\}$ coefficients by K, then the gain K affects neither the stability nor the SNR. In other words, the output of the quantiser will be always the same no matter what is the value of K. Hence, there can be infinite sets of coefficients $\{b_i\}$ to be chosen for a suitable implementation.

3. The designing of the loop filter is carried out by optimisation rather than the traditional filter design method because of the nonlinearity of the system. The SNR is maximised to search for the optimal coefficients $\{a_i\}$ and $\{b_i\}$. The stability is indicated under the reasonable value of SNR.

4. The concept of equivalent quantiser has been introduced to demonstrate the quantisation characteristic of the SDM. It shows that the error characteristic is quite different from the normal uniform quantiser. The width of the division of the input is different from the width of the adjacent quantisation level, i.e., the output points are not the midpoints of the set of the inputs. Furthermore, it is a time-varying quantiser.

5. The maximum possible input magnitudes have been determined for a simple single-loop SDM with sinusoidal input and simulated by computer for more complicated SDM systems. They show that the possible maximum amplitude without causing overload is always smaller than and proportional to the quantisation level.

6. The upper bound of the idle channel noise is proportional to the quantisation

level and inversely proportional to the oversampling ratio.

7. The computer simulations of the SNR for SDM against the bit number, the order of the loop filter, the oversampling ratio, the input signal frequency, and the input magnitude are given to show the properties of the system more intuitively.

# 4

# DISCUSSION OF STABILITY OF ONE-BIT SIGMA-DELTA MODULATION

## 4.1 Introduction

Sigma-Delta Modulation has been known for almost three decades; yet, little has been published comparing experiment with theoretical analysis particularly for nonlinearity and stability of the one-bit system. On the surface this might seem strange because of the apparent simplicity of the sigma-delta system. The principal difficulty of the analysis is the absence of general easy tools for handling stabilities in nonlinear systems with memory. From this viewpoint the simplicity of the sigma-delta modulator is deceptive. This chapter is an attempt to discover more about the important stability aspect in one-bit SDM systems from different angles: nonlinearity, limit cycles, and overload distortion. The chapter is organised as follows. Section 4.2 describes the nonlinearity and presents some of the phenomena. Section 4.3 contains a general discussion of stability issues. Section 4.4 and 4.5 discuss limit cycles which are the important features of nonlinear systems. Section 4.6 addresses the overload problem and its solution.

## 4.2 Nonlinearity in one-bit sigma-delta modulation

Nonlinearities can be classified as smooth nonlinearities and hard nonlinearities [41]. One bit sigma-delta modulators such as one in Fig. 3-3 are systems with hard nonlinearity. A longstanding problem with such nonlinear systems has been the difficulty in analysing their exact behaviour, especially when the nonlinearity is aggravated by its presence in the feedback loop. By far the most common approach is to use a linear approximation in which the coarse quantiser in the feedback loop is replaced by a signal-independent additive white uniform noise source. As a result, the system transfer function of Fig. 3-3 can be separated into two parts: the signal transfer function $F_X(z)$ and the noise transfer function $F_E(z)$. They have been derived in Chapter 2 and are as follows.

$$F_X(z) = \frac{G(z)}{1+G(z)} \quad , \qquad F_E(z) = \frac{1}{1+G(z)} \tag{4-1}$$

From the analysis of Chapter 3, it has been proved that if G(z) is replaced by KG(z), where K is a positive real number, the function of the system will not be affected. However, it is obvious from (4-1) that if G(z) is replaced by KG(z), the functions will change, especially, the noise transfer function. Some researchers tried to describe the stability problem by placing the constraint on $|F_E(z)|$ for $|z|=1$. For example, in [15], $|F_E(e^{j\omega})|$ is described to be less than 2 for the modulator to remain stable. If we replace G(z) in (4-1) by KG(z), $|F_E(e^{j\omega})|$ can be arbitrarily small by choosing K large enough. But the system still operates in exactly the same way. If the

system is originally unstable, it is still unstable even if $|F_E(e^{j\omega})|$ becomes very small. Therefore, using the linear model of (4-1) is not suitable in this situation for the hard nonlinearity of sigma-delta modulators.

In fact, the basic conditions that justify the additive white noise model approximation of the quantiser has been described as follows [27]:

(1) the successive input samples are only moderately correlated;

(2) the number of output levels is large (multi-bit cases);

(3) the output points are very close to the midpoints of the corresponding quantisation intervals.

It is obvious that the one-bit quantiser does not satisfy condition (2). Condition (3) cannot be guaranteed either. That is why the linear model of one-bit sigma-delta modulator may be misleading.

To make the description of the nonlinearity more intuitive, we take the 3rd order SDM system as an illustrated example. The coefficients of the loop filter G(z) are chosen from Table 3-1 as $b_1$=1.0, $b_2$=0.5, $b_3$=0.1301, $a_1$=0.00001, $a_2$=0.00116, and $a_3$=0.0. If , in general case, $\{Kb_i\}$ are used, the function of G(z) will be

$$G(z) = K\frac{(z-1)^2 + 0.5(z-1) + 0.1301}{(z-1)^3 + 0.00001(z-1)^2 + 0.00116(z-1)}$$

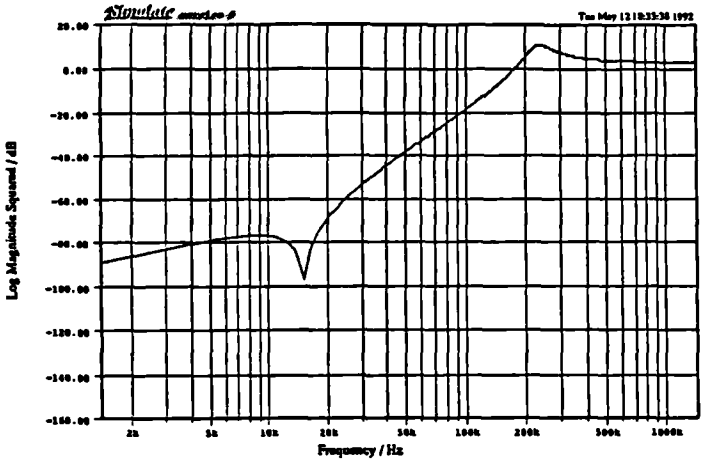
$$= K\frac{z^{-1} - 1.5z^{-2} + 0.6301z^{-3}}{1 - 2.99999\,z^{-1} + 3.00114z^{-2} - 1.00115z^{-3}} \qquad (4\text{-}2)$$

As long as K is a positive real number, the performance of the SDM will remain unchanged. For the model of signal-independent noise, the noise transfer function from (4-1) will be

$$F_E(z) = \frac{1-2.99999z^{-1}+3.00114z^{-2}-1.00115z^{-3}}{1+(K-2.99999)z^{-1}+(3.00114-1.5K)z^{-2}+(0.6301K-1.00115)z^{-3}}$$

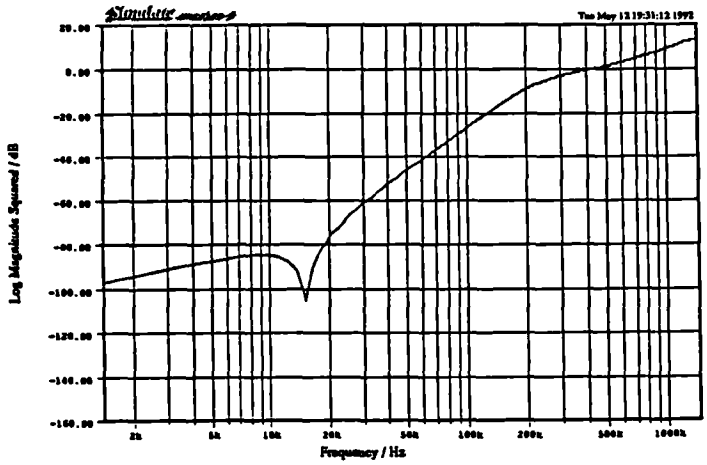
A group of magnitude responses of $F_E(e^{j\omega})$ can be obtained by changing K. Fig. 4-1 gives the magnitude responses of $F_E(e^{j\omega})$ when K=0.8, 2.0, and 2.5. The differences among them are obvious. Also, for some values of K, $F_E(z)$ will be unstable, e.g., when K=0.5 and 3.0, some of the poles are outside the unit circle $|z|=1.0$. However, this group of noise transfer functions map into the same output sequences. It is not a one-to-one mapping. If we define B as a vector space of the coefficients $\{b_i\}$ and Q as a vector space of the output sequences, then we may say that the mapping of B into Q is surjective or this is a mapping of B onto Q [42]. This leads to a conclusion that to make constraints on the magnitude of $F_E(e^{j\omega})$ is not sensible in the design procedure.

(a)



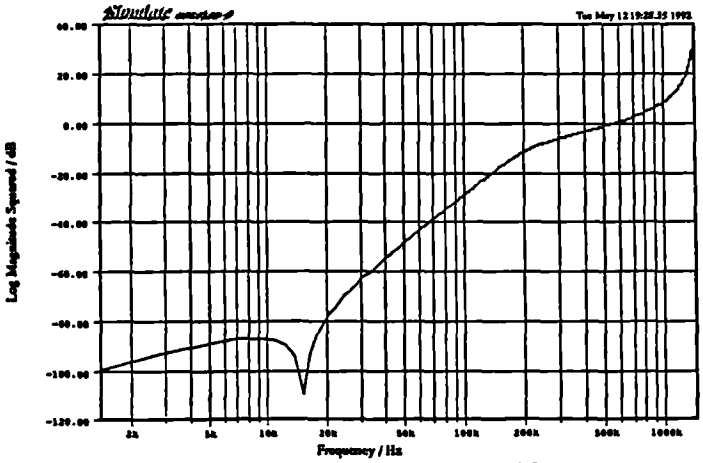

The 3rd order noise shaping function (K=0.8)

(b)




The 3rd order noise shaping function (K=2.0)

(c)




The 3rd order noise shaping function (K=2.5)

Fig. 4-1 Magnitude responses of $F_E(e^{j\omega})$ when K=0.8, 2.0, 2.5

## 4.3 Stability

In spite of the great simplicity of the one-bit sigma-delta modulator as an A/D or a D/A converter, the analyses of its stability are not sufficiently adequate because of the mathematical complexity of the nonlinear feedback loop. In linear systems, instability is equivalent to "blowing up", because unstable poles always lead to exponential growth of the system states. However, for nonlinear systems, blowing up is only one way of instability. Another way is in the form of limit cycles. The stability of a nonlinear system like one-bit SDM, not only depends on the parameters of the system, but also on the initial conditions and the input.

How to describe the stability problem in a SDM system is not very clear so far. The conventional stability concepts may not be used any more. For instance, a one-bit SDM has an output which is bounded by the quantisation level d and hence it is always stable in the sense of Bounded Input Bounded Output (BIBO). Most of the researchers in this field have discussed the stability problem from the aspect of limit cycles. Hein, et al. considered the stability to be a matter of degrees [43]. Stikvoort defined that the system will be considered stable if limit cycles different from the two special types, cannot occur. These two types are : a one-zero pattern at half the sampling frequency in the absence of the signal, and the one caused by a dc input or some offset in the system [31].

In the author's view, the stability problem may be divided into two kinds of situations: the occurrences of overload and/or unwanted limit cycles. The stability problem is strongly related to the overload problem in the SDM systems because of the feedback loop, although overload, in general, does not necessarily imply instability.

For example, in PCM systems, overload can occur but there is no stability problem. In fact, the higher order loop filter in the SDM system amplifies much more the input of the quantiser and makes the quantiser frequently overload, which is reflected by an increase in the amount of quantisation noise. This excess noise is circulated through the loop and can cause an even larger signal to appear at the quantiser input. Perhaps we can say that the sudden drop in SNR curve against input level in Fig. 3-20 when input increases to some level is caused by instability or that it can be considered as the overload distortion of the system. The overload distortion may not lead to the occurrence of the limit cycles, although there may be some relation between them. Limit cycles are evidenced as being essential to the operation of the SDMs [31][6][34]. For example, when the value of a dc input is a rational number, a single loop (1st order) SDM will produce a high frequency limit cycle whose average value will approach dc input. It indicates that we need to use the characteristic of the limit cycles. However, we only need its low-pass version, i.e., its average. If the fundamental frequency and its harmonics with large magnitudes fall into the signal band, the performance of the sigma-delta modulator will be affected. Furthermore, the limit cycles which do not contribute their averages to the approach to the input may be very harmful to the system performance.

Either in overload or limit cycle situation, we may state that the system is unstable if the output can no longer track the input. In the next two sections we will investigate the limit cycles inside the SDM system. Subsequently, in Section 4.6, we investigate the stability from the viewpoint of overload distortion.

## 4.4 Limit cycles in sigma-delta modulation

Limit cycles are unique features of nonlinear systems. In the phase plane, a limit cycle is defined as an isolated closed curve. Trajectories inside the curve and those outside the curve all tend to converge to or diverge from this curve, while a motion started on this curve will stay on it forever, circling periodically around the origin [41]. A limit cycle is a kind of oscillation. However, according to [41], not all the oscillations in the nonlinear system are limit cycles. The stability problem in a sigma-delta modulation system may be partly described as the existence of the unwanted limit cycles or oscillations. A one bit sigma-delta modulator contains the so called "hard" nonlinearity so that in some cases, like dc input, limit cycles will occur [34]. It is generally believed that whether limit cycles will occur or not, the average of the output of the sigma-delta modulator should approximate the input [10].

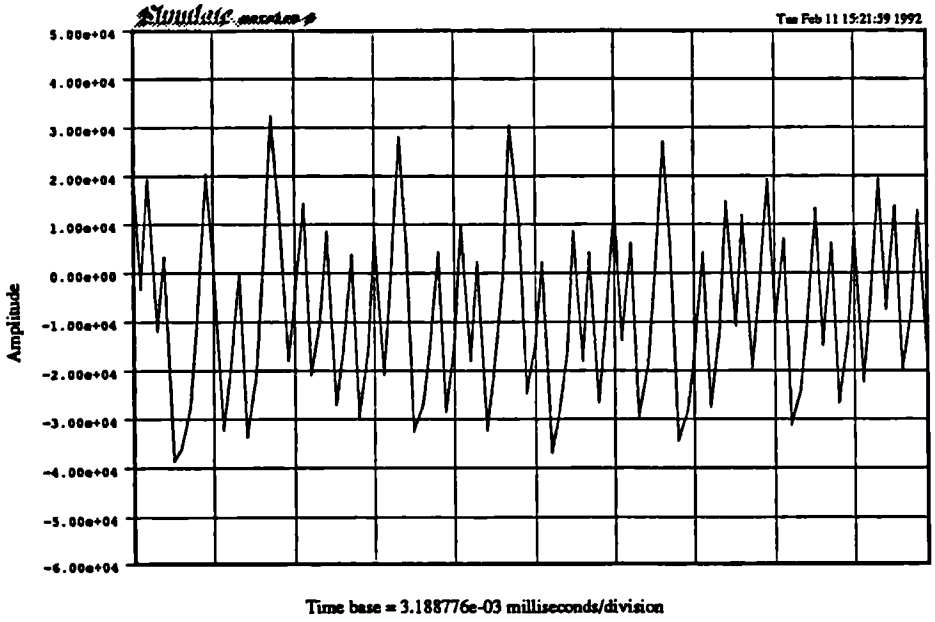
Suppose that an oscillation happens when input is dc so that the output of the loop quantiser q(t) can be expressed as its Fourier series

$$q(t) = e_0 + e_1\cos\omega_0 t + f_1\sin\omega_0 t + e_2\cos 2\omega_0 t + f_2\sin 2\omega_0 t + \cdots \qquad (4\text{-}3)$$

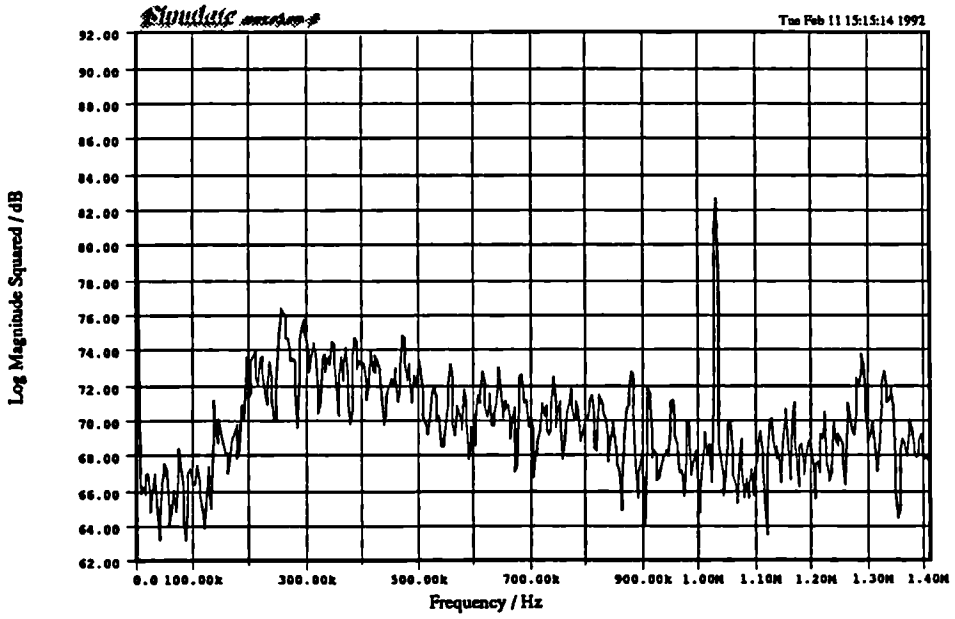
where $e_0$ is its dc component, which should be very close to the value of dc input. In the demodulator, a low-pass filter is needed to remove the unwanted high frequency portion so as to enhance the resolution in the signal band. If some of the frequencies $n\omega_0$, n=1,2, ... are inside the pass band of the low-pass filter, distortion will be introduced. Therefore, it can be seen that sigma-delta modulator uses nonlinearity to produce a kind of limit cycle or oscillation so that its average approaches the input. The

average value (dc) in (4-3) is needed, but if the fundamental frequency and its harmonics with large magnitudes fall into the signal band, the performance of the sigma-delta modulator will be affected. Fig. 4-2 shows an example from the 3rd order sigma-delta modulator with dc input. Fig. 4-2(a) is the time-domain waveform of u(t) before quantiser and Fig. 4-2(b) is its spectrum in the frequency domain. Fig. 4-2(c) is the time-domain waveform of the output q(t) and Fig. 4-2(d) its spectrum. Fig. 4-2(e) is the low-passed version of q(t) and Fig. 4-2(f) its spectrum. It can be seen that in this example, the main limit cycle occurs outside the signal band and the average of q(t) approaches the dc input, where signal band is 0-22.05 kHz and sampling rate is 2822.4 kHz).

It has been shown that the location of the frequency components of the limit cycle varies with the input signal level [34] and thus it is conceivable that for certain input signals, strong frequency components of the noise spectrum will fall into the baseband and hence degrade the signal-to-noise ratio of the system. In the next section, the analysis of the limit cycle for the single-loop (first order) sigma-delta modulator will be given. The exact period can be evaluated if the dc input level is given. Unfortunately, it is very difficult to analyse the higher order system. However, higher order sigma-delta modulation has been shown experimentally to have a less spiky quantisation noise spectrum than does its single-loop counterpart [8].



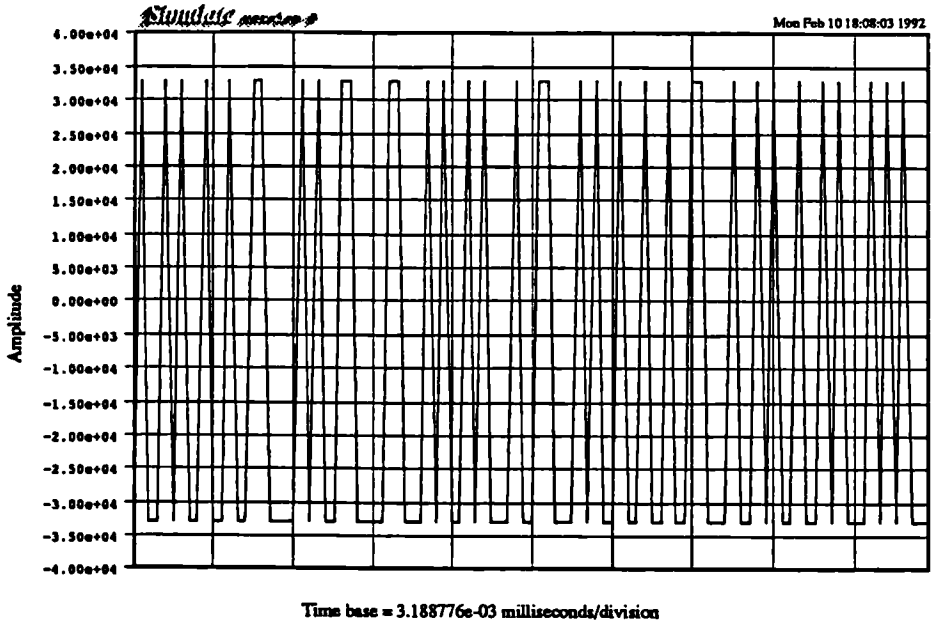

(a)



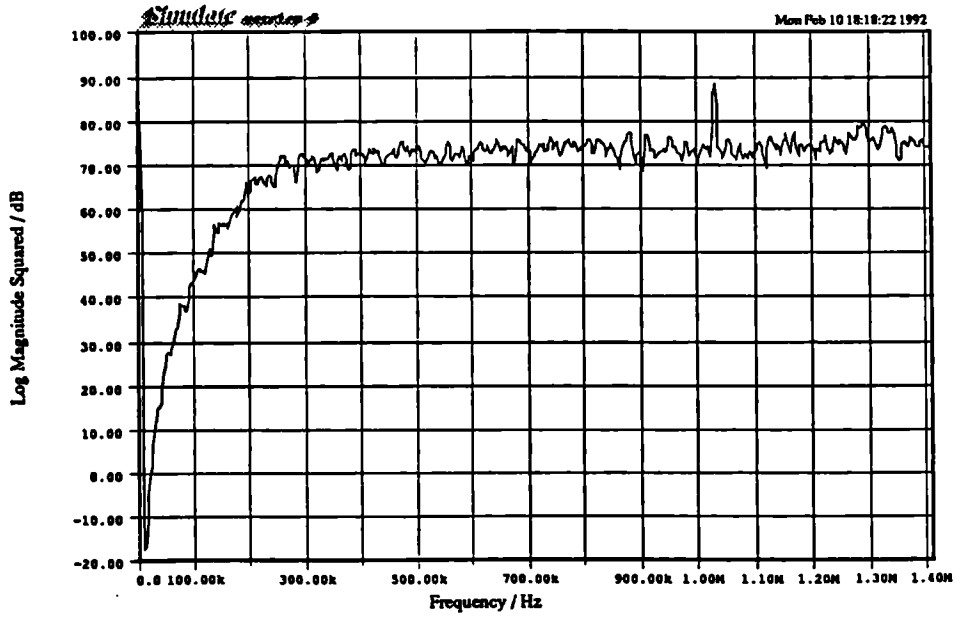

(b)

Fig.4-2 (a) Time-domain waveform of the input of the quantiser u(t)

(b) Spectrum of u(t)



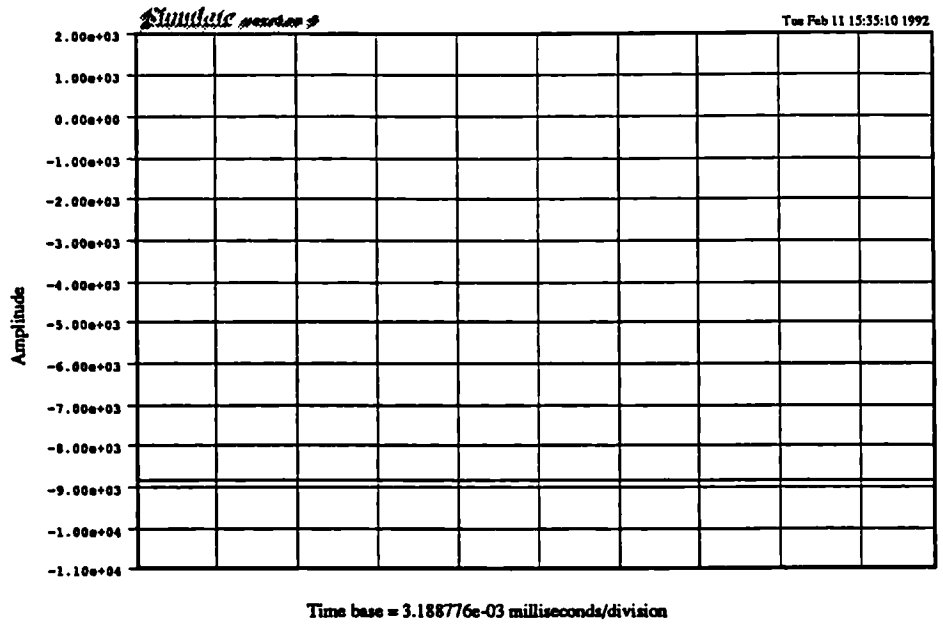

(c)



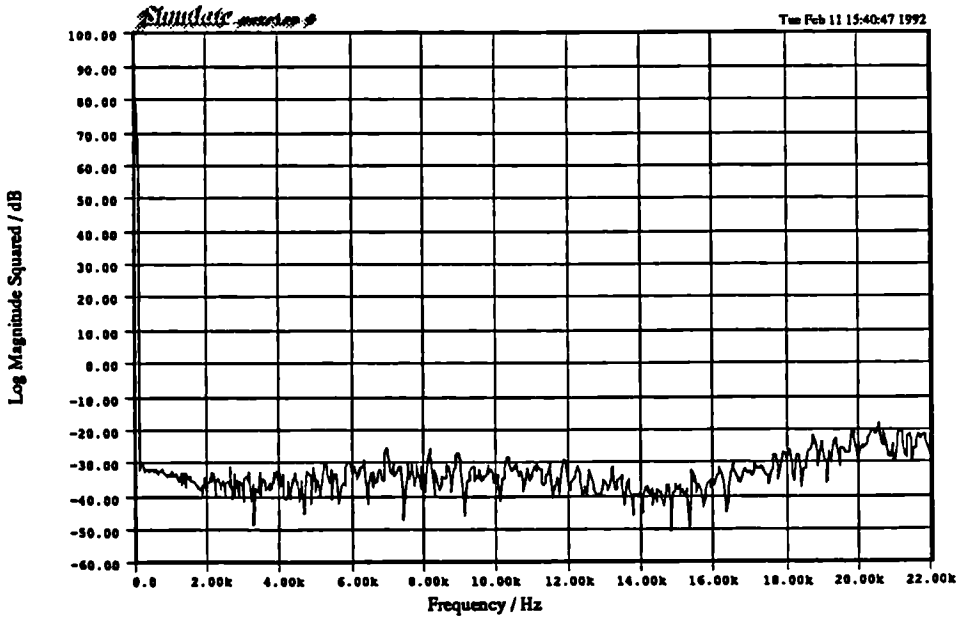

(d)

Fig.4-2 (c) Time-domain waveform of the output of the quantiser q(t)

(d) Spectrum of q(t)




(e)




(f)

Fig.4-2 (e) Time-domain waveform of the output after low-pass filtering

(f) Spectrum of the output after low-pass filtering

## 4.5 Estimating limit cycles of the 1st order sigma-delta modulator with dc input by direct time-domain analysis

As is mentioned in Section 2.8, assuming dc input to the sigma-delta modulator makes analysis much easier and when the oversampling ratio is large or, equivalently, when the input signal freuency is much smaller than the actual sampling rate, the input will appear approximately constant during a short period of time. Thus, all the analyses in this section are based on the condition of dc input.

Supposing that there is a limit cycle of period P in the system described in Fig.3-8(a) with dc input x, if the sum in Fig. 3-8(b) is carried out over P samples, the following will be obtained

$$\frac{1}{P}\sum_{i=0}^{P-1} q_i = x - \frac{u_p - u_0}{P}$$

Because of period P, $u_p=u_0$. Also, assume that there are $k_1$ positive bits and $k_2$ negative bits among the P output bits so that

$$\frac{1}{P}\sum_{i=0}^{P-1} q_i = \frac{k_1 - k_2}{P} d = \frac{b}{a} d = x \qquad (4\text{-}4)$$

where a and b are relatively prime integers, i.e., their greatest common divisor is 1. Assuming that the relationship between the input x and the quantization level d can be expressed as

$$x = \lambda d$$

the necessary condition for the existence of a limit cycle is that $\lambda$ must be a rational number. For the normalised case d being one, the necessary condition is that x must be

a rational number [34]. The length of the limit cycle is a multiple of the denominator a, which can be derived if the dc value of the input is given. In the remaining discussion below in this section, d being one is assumed.

There are two questions which need to be answered:

1) Given a dc input x, which is a rational number, what is the period of the limit cycle;

2) Given a pattern of the limit cycle of the output $q_i$, what is the input value x. They will be answered in reverse order.

Given a pattern of limit cycle with period P, the input value can be easily determined by using (4-4). This is illustrated by the following examples.

*Example 1* The limit cycle pattern of the output is given: +1 +1 +1 -1

so that $k_1=3$, $k_2=1$, and P=4.

Therefore, $x = (3-1)/4 = 0.5$.

*Example 2* The limit cycle pattern of the output is given: +1 -1 -1

so that $k_1=1$, $k_2=2$, and P=3.

Therefore, $x = (1-2)/3 = -1/3$.

Thus, the second question is answered. Now we come to the first.

As we know from (2-32), the upper bound on the absolute error will be smaller as the oversampling ratio becomes larger. If the limit is evaluated as N approaches infinity, the absolute error will approach zero. That is

$$\lim_{N\to\infty} | x - Q_x(x) | = \lim_{N\to\infty} \frac{2}{N} = 0$$

Therefore,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=0}^{N-1}q_i=\lim_{N\to\infty}\left(1-\frac{2N_2}{N}\right)=1-2\lim_{N\to\infty}\frac{N_2}{N}=x \qquad (4\text{-}5)$$

where $N_2$ is the number of negative bits among N bit output. Assume that there are k complete periods among N bit output, that is

$$N = kP + e,\quad N_2 = kk_2 + f,\quad \text{or}\quad N\equiv e\ (\text{mod } k),\quad N_2\equiv f\ (\text{mod } k)$$

where e and f are the reminders, then (4-5) will become

$$x = 1 - 2\lim_{k\to\infty}\frac{kk_2+f}{kP+e} = 1-\frac{2k_2}{P} \qquad (4\text{-}6)$$

It can be proved that $k_2$ and P are relatively prime (see Appendix C). Therefore, given the value of dc input x which can be expressed as a rational number, then the period P can be determined through the equation derived from (4-6)

$$\frac{k_2}{P}=\frac{1-x}{2}$$

where $k_2$ and P are relatively prime.

*Example 3* $x = 0.5$

$k_2/P = (1-0.5)/2 = 1/4$ so that $P = 4$

which is consistent with the result in Table 2-2.

*Example 4* $x = -0.65$

$k_2/P = (1+0.65)/2 = 33/40$ so that $P=40$

The frequency which is corresponding to P will be $f = f_s/P$, while $f_s$ is the sampling frequency.

In the case of first order SDM, almost all dc inputs give rise to a limit cycle at the quantiser output. That is, if x is chosen from a uniform distribution within (-d, d), the probability that this input gives rise to a limit cycle is 1. Furthermore, only one limit cycle exists for a fixed x [44]. The period of the limit cycle P can be decided by (4-6) and is independent of the initial condition. However, this is not the case for higher order system such as double loop SDM. The existance of the limit cycle and its period not only depends on the input but also the initial conditions [34].

The limit cycles produced by the single loop SDM have the property that the ones and minus ones are distributed as uniformly as possible in the output stream [34]. This indicates that the limit cycles tend to have a high fundamental frequency which will be outside the signal-band with high probability.

## 4.6 Overload and the use of clippers

In this section, we will investigate the stability problem from the overload point of view. The question is why the conventional noise transfer function $F_E(z) = (1-z^{-1})^n$ when n is greater than 2 will cause the system to be unstable, while the optimised high order filters will not do so. In order to understand the reason, it is necessary to observe the spectrum of the error of quantiser which is inside the SDM loop. Fig. 4-3 gives a diagram for calculating this error called e(k). The input x(k) is a sinusoidal signal with the frequency of 10087 Hz. Fig. 4-4 shows the error spectrum when using the conventional third order noise transfer function $F_E(z) = (1-z^{-1})^3$. It is observed that the error spectrum is not white at all. Fig. 4-5 gives the error spectrum when G(z) from (4-2) is used. The error in this case is more like white noise compared with Fig. 4-4.

This may be explained from the angle of overload. As is known, the granular noise tends to be white while the overload distortion will be more spiky and more fluctuated. The SDM system based on the conventional noise transfer function frequently causes overload of the quantiser. Relatively, the optimised filter will alleviate the overload distortion. Therefore, the error in the optimised system tends to be more granular noise and looks more like white noise.

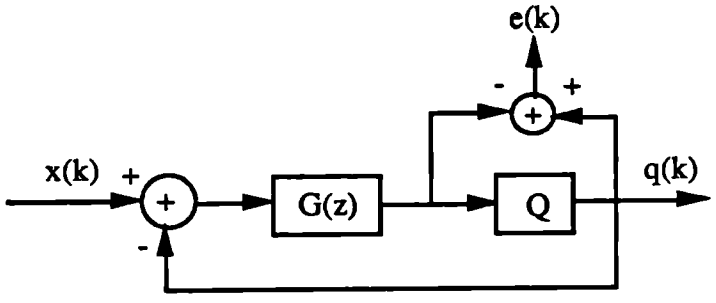



Fig. 4-3 Diagram of calculating quantiser error

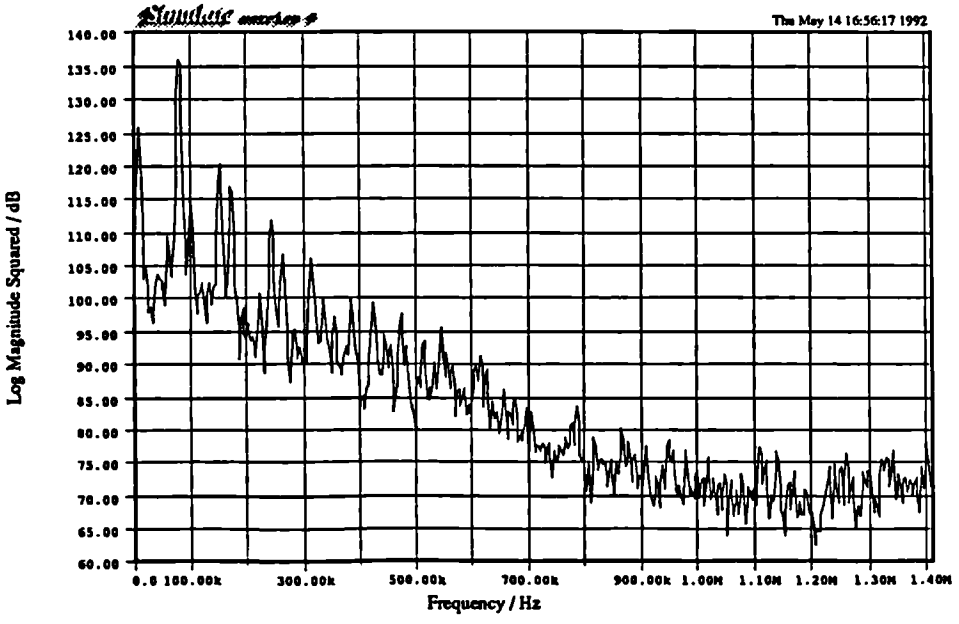



Fig. 4-4 Error spectrum when using the conventional noise transfer function $(1-z^{-1})^3$

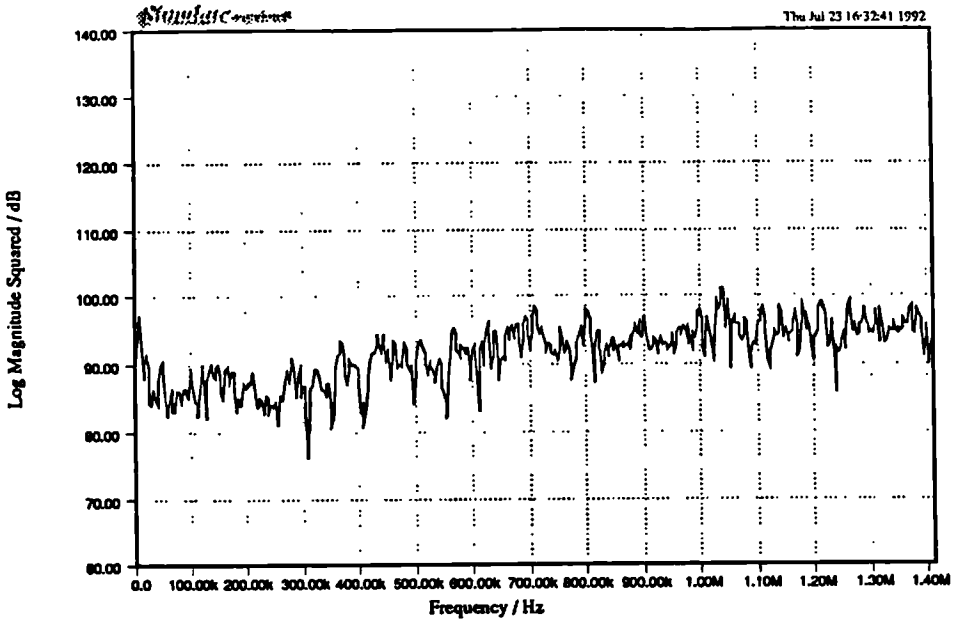



Fig. 4-5 Error spectrum when using the optimised 3rd order filter

The one-bit sigma-delta modulator is a nonlinear system. The stability of this kind of systems, as mentioned in the previous section, not only depends on the parameters of the system, but also the initial conditions and the input. It is observed that under the same conditions, i.e., the same sets of parameters and the same initial states of the system, the 3rd order system can be stable with the certain sinusoidal inputs without adding clippers, but oscillations will sometimes appear when the input is the 15-second piece of music. Fig. 4-6 shows a short period of time-domain waveform together with the signal waveform after being processed, i.e., the reconstructed signal. It is clearly shown that the reconstructed signal contains oscillations. These kinds of phenomena are observed as being caused by overload. In order to solve this problem,

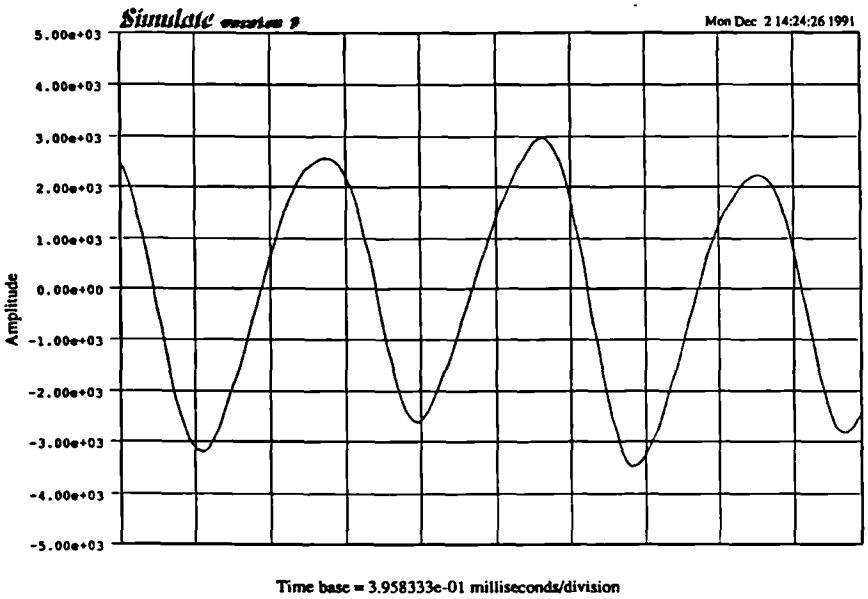
clippers are needed to saturate the magnitude deliberately. If one clipper is placed just before the quantiser, it will not play any role because one-bit quantiser itself has the function of saturation. Considering each integrator in Fig. 3-3 whose transfer function is $z^{-1}/(1-z^{-1})$, its magnitude response will be $(2-2\cos\omega)^{-1/2}$. It has enormous gain at low frequencies near dc. It seems reasonable to place a clipper after each integrator, as is shown in Fig. 4-7. A simple clipping model is used which employs a saturation characteristic of the form

$$\text{sat}(x) = \begin{cases} x, & \text{for } |x| \le M \\ M\,\text{sign}(x), & \text{otherwise} \end{cases} \tag{4-7}$$

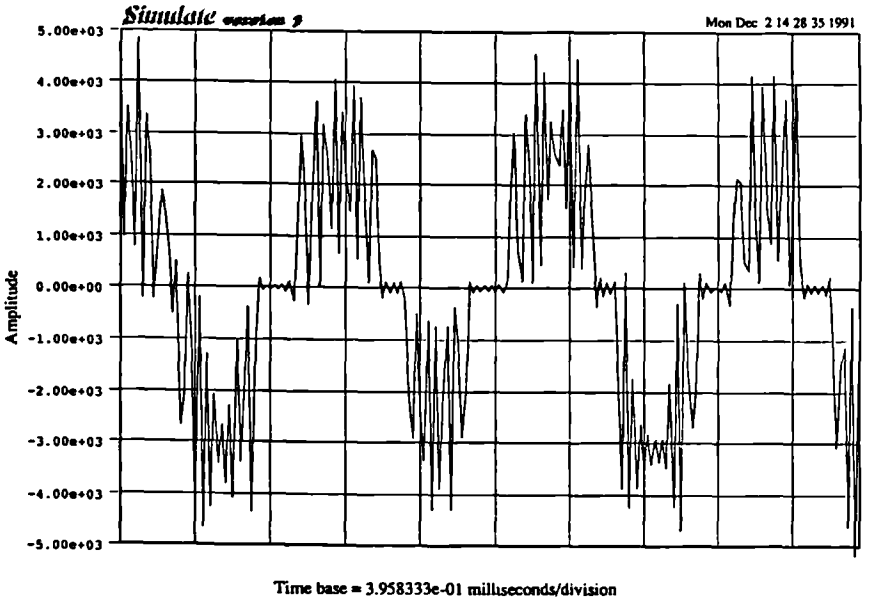
where M is the clipping level, and sign(·) is the signum function. The clippers are typically designed so that the clipping level is not much larger than the quantisation level d. For each stage of integrator in Fig. 4-7, the clipping level is different because the signal level becomes higher and higher as the stage number increases. The following levels for each stage is based on the computer simulations

$$M_j = m\,(1{\cdot}2)^{j-1}\,d \qquad (j=1, 2, 3, 4)$$

where $M_j$ represents the clipping level of the clipper j, d is the quantisation level, and m is a factor which can be adjusted between 4 and 10 depending on the coefficients of the loop filter G(z). After adding the clippers, the phenomenon in Fig. 4-6 disappears. It shows that the presence of the clippers in the SDM strongly influences the stability of the third or higher order systems.




(a) Original signal




(b) Reconstructed signal

Fig.4-6 Oscillation occurs in the music signal case

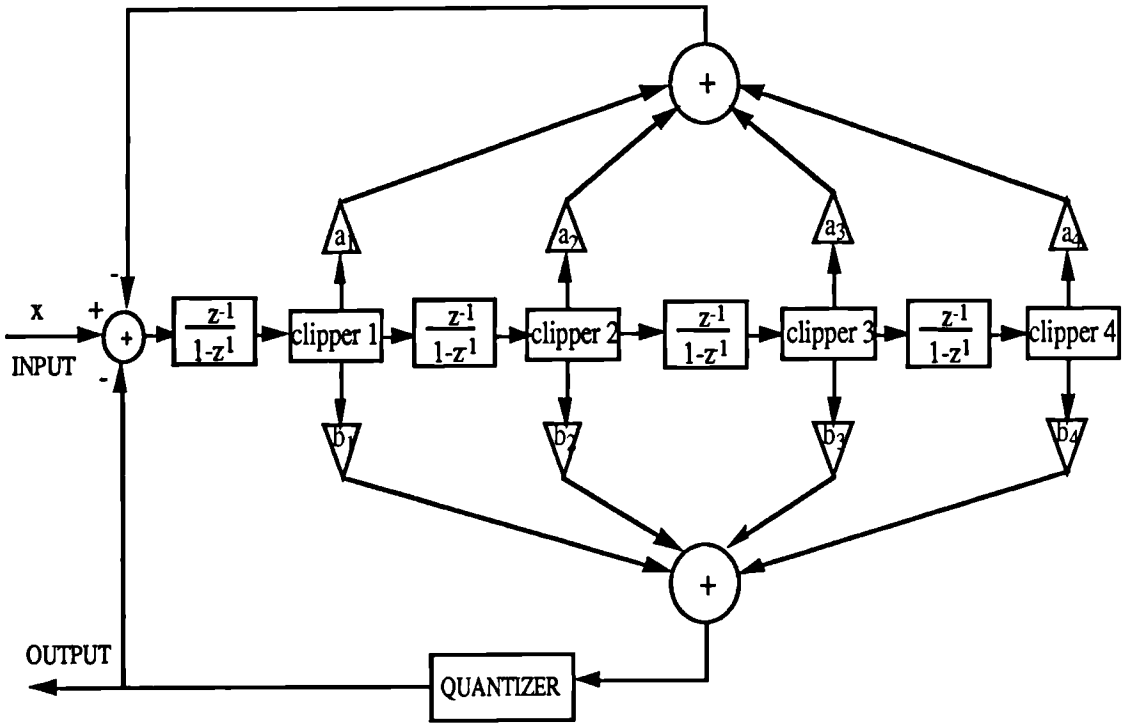



Fig. 4-7 A fouth order SDM with clippers

Saturation or clipping implies loss of state information and performance degradation. However, if the magnitude of the signals of each stage grows more and more and hence the severe overload situation dominates the system, we will have even worse degradation.

## 4.7 Summary

Sigma-delta modulators are nonlinear systems. It has been demonstrated that the traditional filter design methods for the noise shaping function are not suitable for the design of the loop filter. This is the reason for choosing the optimisation method in the design procedure in Chapter 3.

The stability is still a difficult topic for a sigma-delta modulation system. The system is always stable in the sense of BIBO. Its stability is reflected in overload and limit cycle situations in which the output of the system may no longer track the input.

Some limit cycles are essential to produce the oscillation so that its average will approach the input, but some of them are harmful. The period of the limit cycle produced by a simple first order SDM can be determined when a dc input is applied. However, for more complicated systems, it is more difficult to calculate the exact period of the limit cycles.

Overload distortion can be alleviated by using one clipper after each stage of integrator. The clipping levels are different for each stage and become higher as the stage number increases to accommodate the increased magnitude of the signal.

# 5

# ADAPTIVE QUANTISER FOR SIGMA-DELTA MODULATION

## 5.1 Introduction

The advantages of sigma-delta A/D and D/A converters over traditional PCM are by now well known. One-bit coding offers attractive possibilities in A/D and D/A conversion for audio, which is, in part, due to the fact that it does not require precise component matching. However, a one-bit system usually needs either a very high oversampling ratio, or a very high order loop filter, in order to achieve the audio quality. For example, in a one-bit third order system, to obtain the quality equivalent to 16 bit PCM, the oversampling ratio should be at least 128. The higher the oversampling ratio, the more difficult the implementation. Also, the higher the order of loop filter, the more likely the system will be unstable, and generally, the more difficult it is to design.

As is well known, the quantiser also plays a key role in a SDM besides the oversampling ratio and noise shaping function. However, little work has been carried

out on the quantiser itself. Usually, quantisers with fixed step size are used for sigma-delta modulators and these cannot always match the variance of the input. For small signal magnitude, very coarse quantisation could happen and for large magnitude of signal, overload may occur. Adaptive quantisers have been used in some digital coding systems such as APCM, ADPCM, ADM [29]. They have also been examined in some simple sigma-delta modulators before [45] [46]. They have shown great advantages in increasing the dynamic range and attaining better quality at the same bit rate or reducing the bit rate of the system while maintaining the same quality.

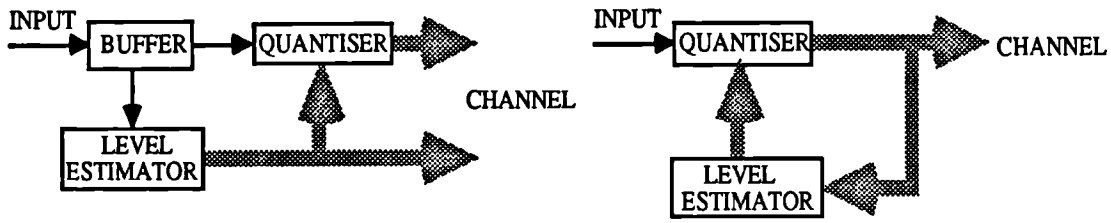
The main topic of this chapter is to investigate adaptive quantiser with digital logic in sigma-delta modulation. Basic methods of adaptation and the design of the adaptation logic are described in Sections 5.2 and 5.3. The adaptive SDM system is then simulated for both sinusoidal and music signals in Sections 5.4 and 5.6. Section 5.5 shows some results with an adaptive SDM working as an A/D converter. Furthermore, the quantisation effect on adaptation levels is discussed and the idle channel noise is shown to compare with that in the fixed SDM system. Finally, Section 5.10 gives the conclusions.

## 5.2 Adaptive quantisation

The magnitude of a music signal can vary over a wide range depending on the instruments, singers, etc. In the digitising process, on the one hand we wish to choose the quantisation step size large enough to accommodate the maximum peak-to-peak range of the signal; on the other hand we would like to make the quantisation step small so as to minimise the quantisation noise and the idle channel

noise. To satisfy both of them is impossible by using a fixed quantiser. The basic idea of adaptive quantisation is to let the step size vary so as to match the variance of the input signal.

In order to adapt the step size, it is necessary to obtain an estimate of the time varying amplitude properties of the input signal. Usually, there are two kinds of methods: feed-forward and feedback adaptation[28]. Feed-forward adaptation is based on the estimation of unquantised samples, i.e., usually at the input of the quantiser. Feedback adaptation is based on the estimation of the output of the quantiser. Their block diagrams appear in Fig. 5-1. Feed-forward estimates of step-size are unaffected by quantisation noise: therefore, they are more reliable. However, the system needs to transmit this additional information to the receiver. Although the feedback estimates are not as accurate as the feed-forward ones, there are no additional bits needed for the estimation.




(a) Feed-forward adaptation (b) Feedback adaptation

Fig. 5-1 Feed-forward and feedback adaptation

A common approach to variance calculation is to assume that the variance is proportional to the short-time energy, which is defined as the output of a low-pass filter with the squared signal as its input, $x^2(k)$ [29]. That is

$$\sigma_x^2(i) = \sum_k x^2(k)\, h(i-k)$$

where h(k) is the impulse response of the low-pass filter. If a one-bit quantisation function is defined by

$$q(k) = \begin{cases} d, & \text{if } u(k) \geq 0 \\ -d, & \text{otherwise} \end{cases}$$

where u(k) is the input of the quantiser, then the adaptive logic can be

Feed-forward: $d_i = c\, \sigma_u^2(i)$

Feedback: $d_i = c\, \sigma_q^2(i)$

where c is a scaling constant, and $\sigma_u^2$ and $\sigma_q^2$ are the variances of the signal u(k) and q(k) respectively. It is obvious that the feedback logic is always constant so that it cannot be used for the one-bit case.

To obtain the short-time energy needs a large amount of calculation. An alternative approach is to use local values of peak output magnitude to vary the overload level. Thus for a feed-forward adaptation of one-bit quantiser, the logic can be in the form

$$d_i = c\, |u|_{max}; \qquad |u|_{max} = \max \{ |u(i-k)| \}; \quad k=1,2,...,K$$

where $d_i$ is the one-bit quantisation level of the ith data block and each block consists of K samples. Feedback adaptation based on the maximum magnitude of the output of a one-bit quantiser is also meaningless because the maximum is always the same as the

quantisation level. The maximum-magnitude logic is simpler than the variance estimation and particularly appropriate to the control of overload distortion.

Considering the case of sigma-delta modulation, the additional bits to be stored or transmitted are unwanted. In particular, if SDM is used as an ADC, then the signal before the quantiser is analogue. We would like to use digital logic to reduce the complexity of the circuitry. Hence, feedback adaptation has to be chosen. However, feedback adaptation based on the output of the quantiser is impossible due to the characteristic of single-bit quantiser. Therefore, a feedback logic which is not directly based on the output of the quantiser has to be introduced. An estimate of the maximum magnitude of the input is used for the adaptation and will be described in detail in the next section.

## 5.3 Logic design of adaptation for SDM

As we have seen in Chapters 2 and 3, the upper bound of the quantisation error and the idle channel noise is proportional to the quantisation level d (see 2.8 and 3.6). The maximum input level also depends on d. We are confronted with a dilemma in quantising the signal. On the one hand, we would like to choose small d to reduce the quantisation error and idle channel noise. On the other hand, d has to be large enough to prevent the overload distortion. In this section, we will discuss how to design the adaptation logic for a SDM system.

As is mentioned in Section 5.2, adaptive quantisation based on the maximum-magnitude logic is easier than that based on the variance estimation. And also, it is more suitable to SDM systems which are very sensitive to overload distortion. If the

step size of the quantiser can be adjusted as the maximum magnitude of the input u(k) to the quantiser changes, the dynamic range of the sigma-delta modulator can be increased. However, it is difficult to estimate the magnitude of u(k) from q(k) because of the coarse quantisation: in the case of one-bit quantiser, q(k) just represents the sign of u(k). But the quantisation level must have some relation with x(k). As was described in Section 3.4, the sigma-delta modulator-demodulator can be considered as an equivalent quantiser. Based on this concept, its quantisation level may be changed by using feedback adaptation according to its output. Fig. 5-2 shows the basic idea. Supposing that over N samples, input x is relatively constant, from equation (3-1) we know that

$$Q_x(x) = \frac{k_1 - k_2}{N} d \tag{5-1}$$

for the single-loop SDM, where N is the oversampling ratio, $k_1$ and $k_2$ are positive or zero and negative samples over N samples respectively. Equation (5-1) indicates that changing the quantisation level of the equivalent quantiser can be carried out by equivalently changing the quantisation level d of the real quantiser.

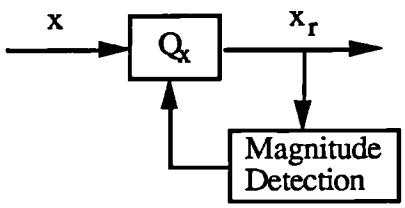



Fig. 5-2 Adaptive quantisation based on the concept of equivalent quantiser

The computer simulations have shown that when the input level is constant, there is an optimal quantisation level $d_{opt}$ by which the maximum signal-to-noise ratio can be reached. The characteristic graphs are illustrated in Fig. 5-3. The level d being smaller

than $d_{opt}$ corresponds to overload situation whereas when d is greater than $d_{opt}$, the quantisation noise increases.

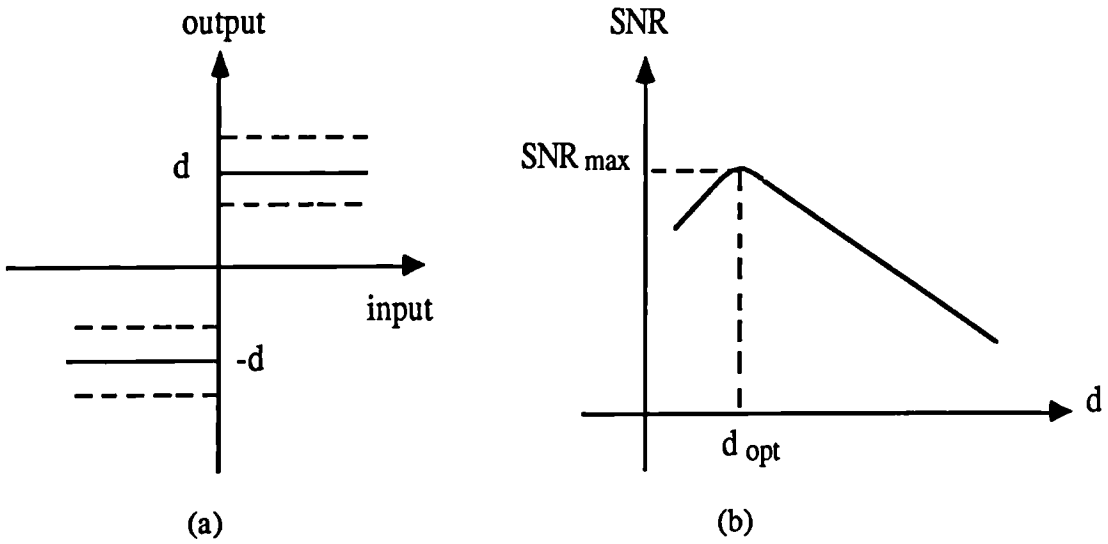



Fig. 5-3 Characteristic graphs of changing the quantisation level: (a) The characteristic of the quantiser; (b) The signal-to-noise ratio against the quantisation level when input magnitude is a constant

The magnitude of the output $x_r$ can be detected by first low-pass filtering the output q(k) of the one-bit quantiser and then detecting the maximum magnitude over a certain period of time. According to the maximum-magnitude logic, it is reasonable to have the adaptive logic as follows

$$d_{i+1} = c\, M_i, \quad d_{min} \leq d_{i+1} \leq d_{max} \tag{5-2}$$

where $M_i$ is the maximal magnitude in the ith block of the output samples of the low-pass filter, that is, the maximum magnitude estimate of the input x(k) for the

ith block, and $d_{i+1}$ is the step size for the (i+1)th block of samples. The factor c is a constant which depends on the order of the system and the oversampling ratio. The optimal value of it for sinusoidal input is $C_{opt}$ which has been derived and computer-simulated in Chapter 3. Each block contains, say, K samples and K depends on the stationary property of the signal. For speech, usually the signal can be considered stationary over 10-30 ms periods. For music signals, it may be less than 5 ms, that is, K should be less than 220, if the sampling frequency is 44.1 kHz. For both modulator and demodulator, an appropriate initial value $M_0$ is required. In (5-2), $d_{max}$ depends on the maximum level of the system. $d_{min}$ depends on the requirement for noise level when the input is zero.

If the SNR for an A/D system with no oversampling and noise-shaping is SNR= $(\sigma_x^2/\sigma_e^2)$, where $\sigma_x^2$ is the signal power and $\sigma_e^2$ is the quantisation noise power, the total SNR for a sigma-delta modulator is

$$SNR = (\sigma_x^2/\sigma_e^2)\, SNR_{enhancement}$$

where $SNR_{enhancement}$ is obtained from oversampling and noise shaping techniques. For a sinusoidal input: $x = E\sin\omega_0 t$, $\sigma_x^2 = 0.5E^2$. In the case of no overload, and assuming that the quantisation noise is uniformly distributed in the range $|e| \le d$, with the probability density function $p_e(e)$ being $1/(2d)$, then

$$\sigma_e^2 = \int_{-d}^{d} e^2\, p_e(e)\, de = \int_{-d}^{d} \frac{e^2}{2d}\, de = \frac{d^2}{3}$$

Thus

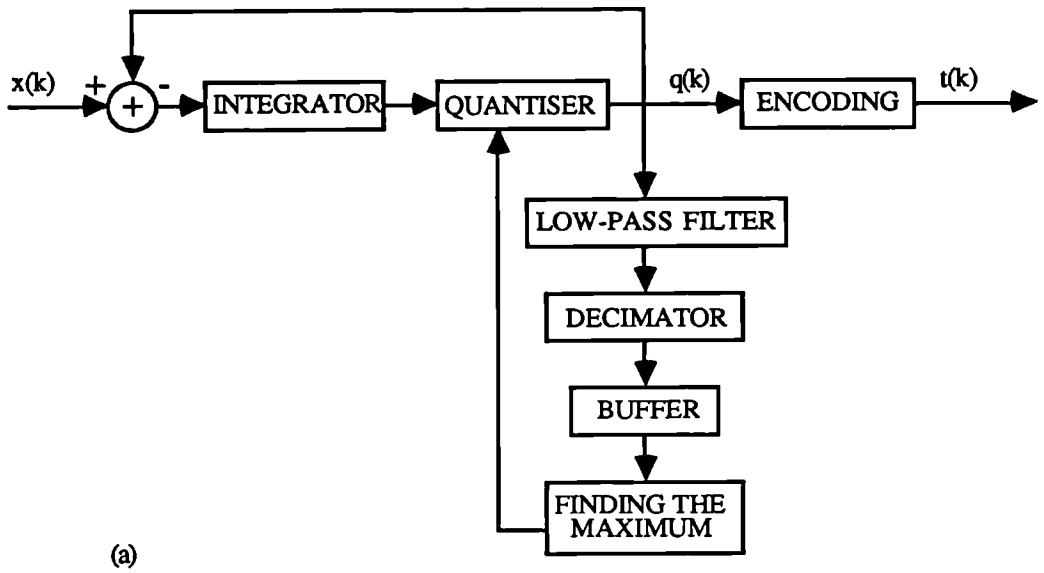
$$SNR = ( 3E^2 /2d^2 ) SNR_{enhancement} \quad (5\text{-}3)$$

Assuming the ideal case: $d_i = cE$ according to equation (5-2), then

$$SNR = ( 3/2c^2 ) SNR_{enhancement} \quad (5\text{-}4)$$

which means that SNR can be independent of the input level. This deduction is based on the model of additive white noise. It needs to be tested by computer simulation because, strictly speaking, a one-bit SDM is a nonlinear system. From (5-4) it can be seen that the smaller the c is, the better SNR can be obtained as long as it does not cause overload. As was mentioned before, the value of c represents the relation between the input level and the quantisation level d. Its optimal values for different systems with the sinusoidal input have been calculated and simulated in Section 3.5.

The diagram of a discrete-time model of adaptive sigma-delta modulator is shown in Fig. 5-4(a) and the corresponding demodulator is shown in Fig. 5-4(b), where t(k) is a digital sequence which is either 1 or 0 and q(k) is an analogue sequence after decoding whose value changes according to the adaptation logic. Both modulator and demodulator use the same kind of low-pass filter and decimator. After the decimator, samples are stored in a buffer with length K. Then the maximum value is detected among K samples. The main purpose for using low-pass filtering and decimation in the demodulator is to obtain high quality for the reconstructed signal. Simultaneously, the estimate of the magnitude of x(k) can be obtained. But in the modulator, the only purpose is to find the maximum magnitude of the signal so that a very sharp low-pass filter is not necessary. Thus a simple low-pass filter can be used in the modulator while the complexity of the demodulator will be greater because two sets of low-pass and decimation systems have to be used.



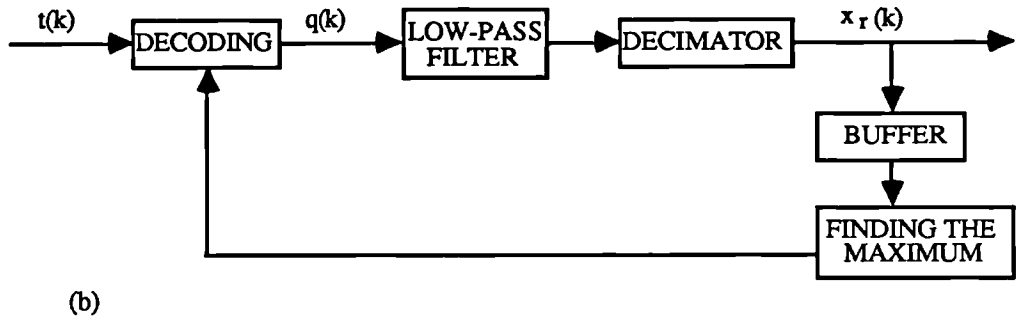




Fig. 5-4 Adaptive SDM (a) modulator; (b) demodulator

## 5.4 Signal-to-noise ratio tests and multi-tone test

Adaptive quantisation has been carried out for one-bit, 1st, 2nd, and 3rd order sigma-delta modulators. The structure in Fig. 3-3 is used for the simulations. Considering that for a music signal, the stationary time may be less than 5 ms, and the controlling factor $M_i$ for the current block of samples is calculated from the previous block, K=60 is chosen as the block size, which corresponds to 1.36 ms. This K value could be changed according to the statistical characteristics of the input, which will be discussed later. For fixed quantisers, assuming that M is the maximum input level which does not cause overload, cM is chosen as the quantisation level. The input x(k) to the modulator and the output $x_r(k)$ of the demodulator are discrete time analogue signals, i.e., in computer simulation, floating point numbers are used to represent them.

The SNR results of both fixed and adaptive 3rd order sigma-delta modulation, as the input level changes from 0 dB (maximum) to -60 dB, are shown in Fig. 5-5, which clearly shows that the dynamic range of the system can be improved effectively by using an adaptive quantiser. Fig. 5-6 gives the spectrum results when the input level is -60 dB. Fig. 5-6(a) is the spectrum of the reconstructed signal when using a fixed quantiser and Fig. 5-6(b), when an adaptive quantiser is applied. It can be seen that the noise level is at -180 dB with respect to the full scale for the adaptive system. Fig. 5-5 also gives the results of the 1st and 2nd order adaptive sigma-delta modulators. They all have the same property: SNR is nearly independent of input level, which is consistent with the result from equation (5-4). Fig. 5-7 gives the results of fixed and adaptive 3rd order sigma-delta modulators when the input contains three tones whose total level is +10 dB which represents the overload case for a fixed

SDM. Fig. 5-7(a) shows the spectrum of the reconstructed signal of the fixed SDM, from which it can be seen that because of overload, the effects of harmonic and intermodulated components are very severe. However, in Fig. 5-7(b), the harmonic and intermodulation distortion is reduced effectively by allowing the quantisation level to increase as the input magnitude becomes larger.

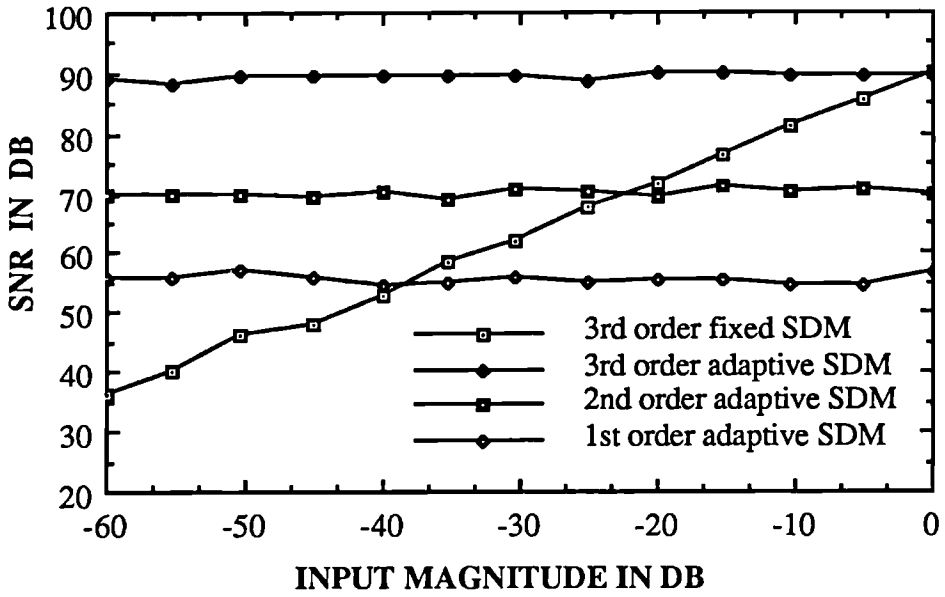
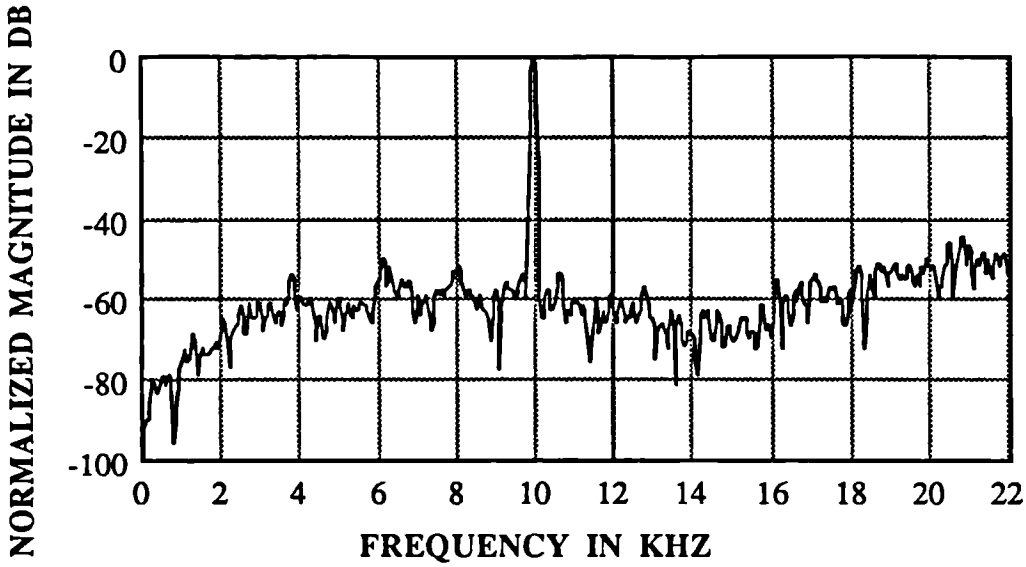
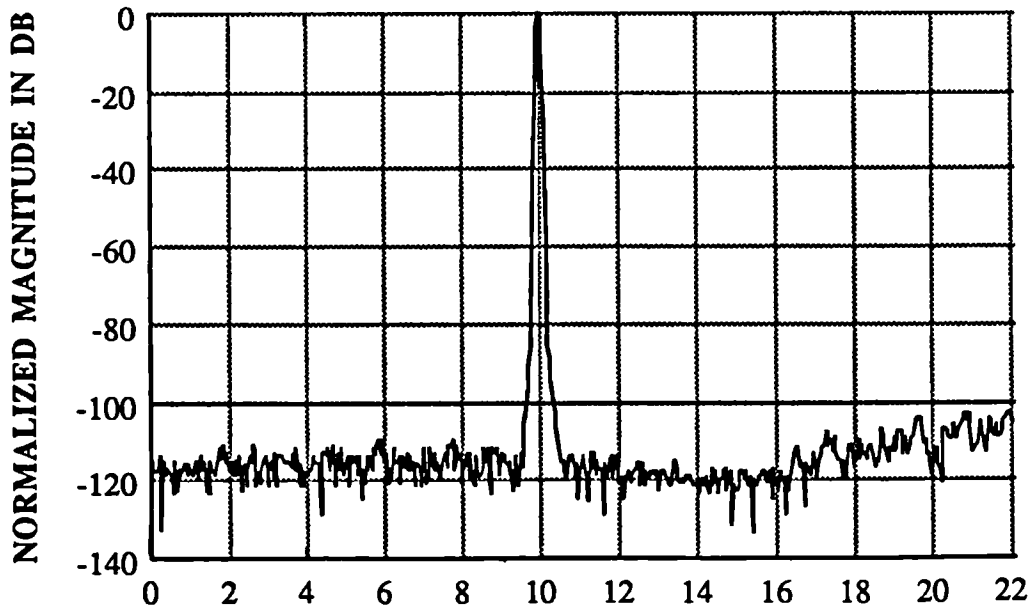



Fig. 5-5 SNR results for the fixed and adaptive 3rd order sigma-delta modulators, the 1st and 2nd order adaptive sigma-delta modulators ( input: 10087 Hz sinewave; oversampling ratio: 64; Nyquist sampling frequency: 44.1 kHz; 1-bit quantiser. )

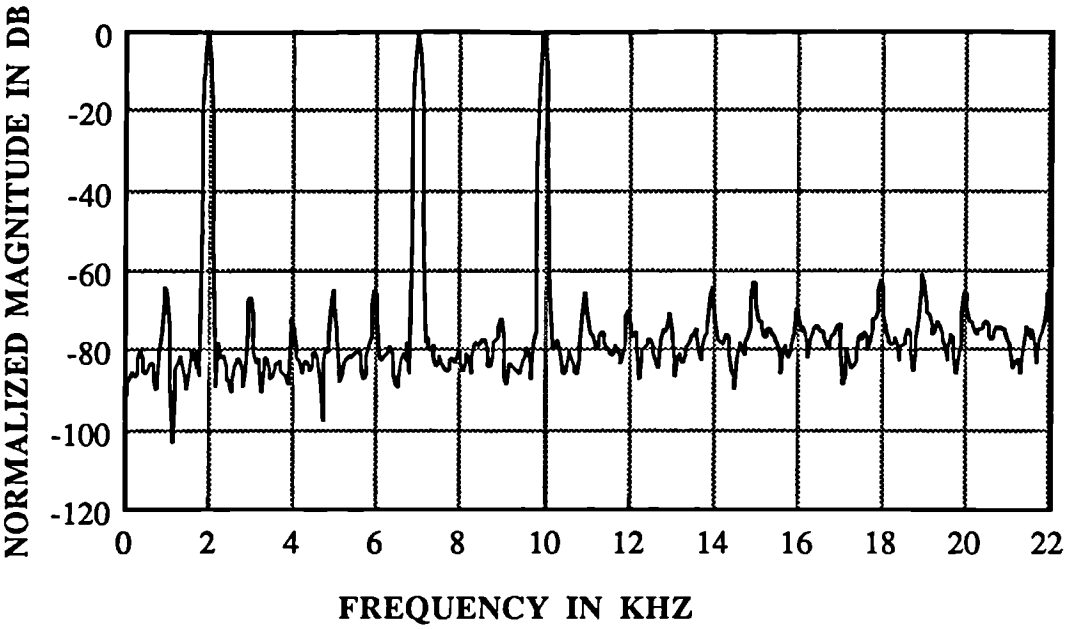



(a) Using fixed quantiser

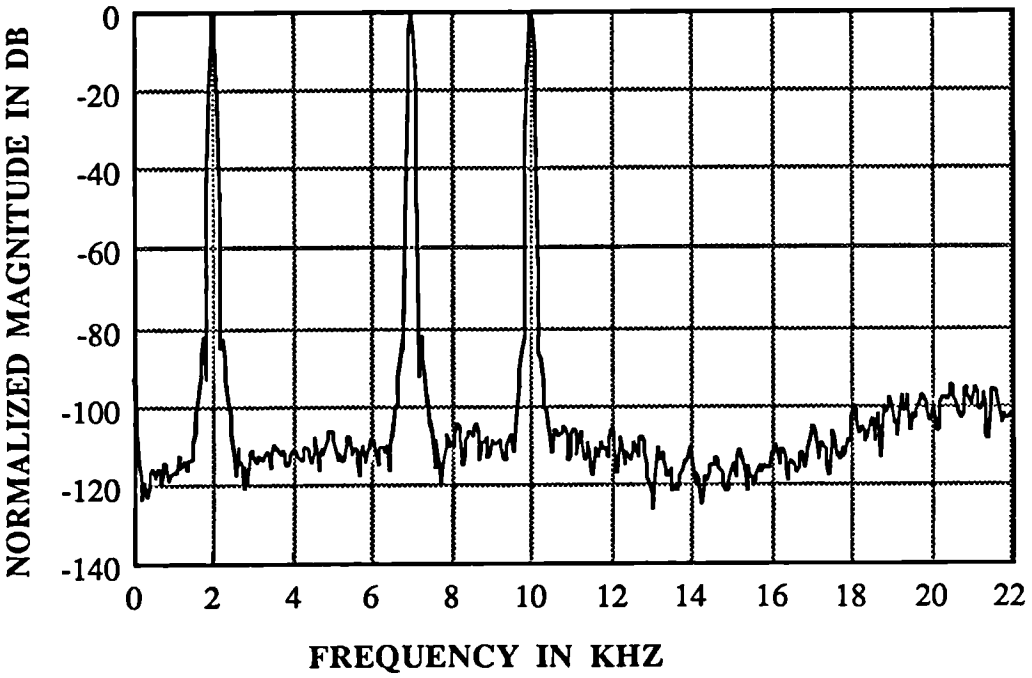



(b) Using adaptive quantiser

Fig. 5-6 Comparison of spectra of reconstructed signals between 3rd order fixed and adaptive SDMs when input level is -60 dB




(a) Using fixed quantiser




(b) Using adaptive quantiser

Fig. 5-7 Comparison of spectra of reconstructed signals between 3rd order fixed and adaptive SDMs when input contains 3 tones and the total input level is 10 dB

Fig. 5-8 gives an illustrating graph which is an SNR comparison between an adaptive 3rd order SDM (oversampling ratio: 64) and linear PCMs. It shows that as the input level decreases, the SNR of the adaptive SDM exceeds that of higher and higher bit PCM systems. After -33 dB, it is even better than a 20 bit PCM system.

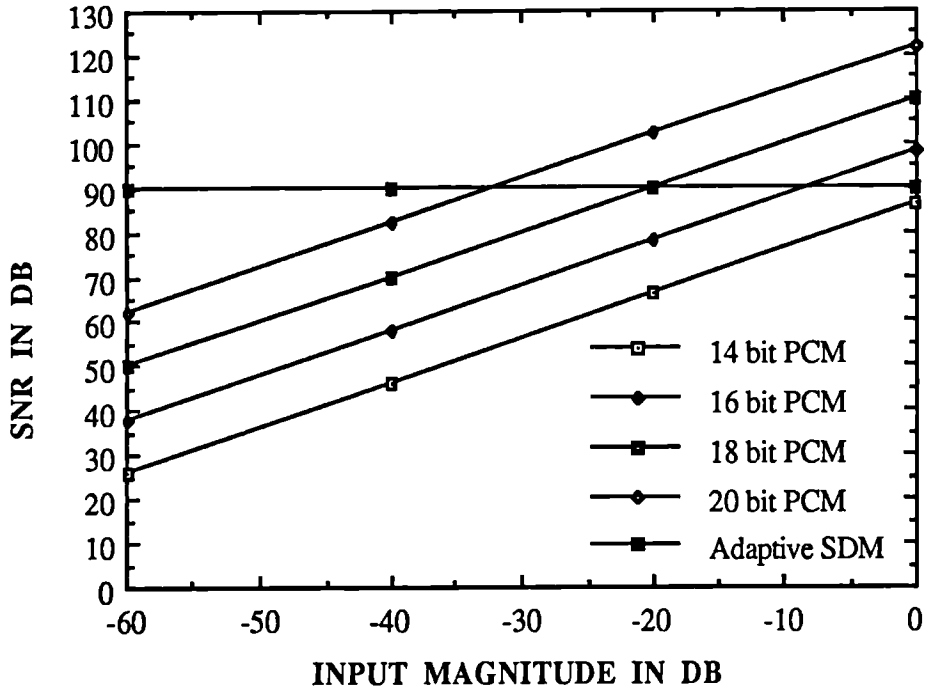



Fig. 5-8 Comparison between 3rd order adaptive SDM with oversampling ratio 64 (average value) and linear PCMs (theoretical values)

## 5.5 Adaptive sigma-delta modulator used as an A/D or a D/A converter

The diagram of an A/D converter using sigma-delta modulator is shown in Fig. 5-9. It can be seen that the differences between a fixed and an adaptive sigma-delta A/D converter are the logic block for finding the maximum and an extra D/A converter in Fig. 5-9. It only needs a buffer and simple calculations to find the maximum, but a D/A converter indicates the increase of the complexity. We will discuss the possible implementation in Section 5.8. When PCM rounding is added after the decimator, the final quality depends on both sigma-delta modulator and N bit linear PCM. It cannot be better than either of them.

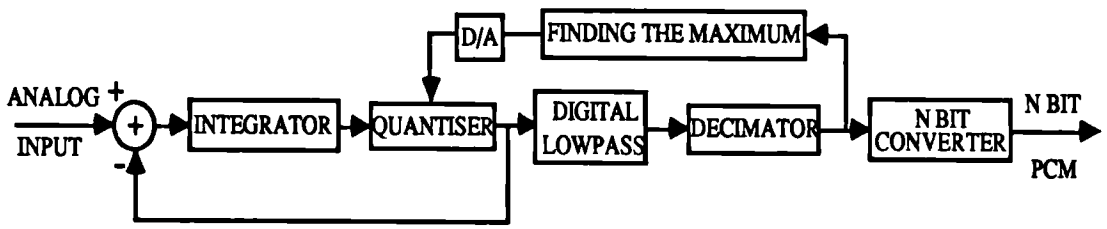



Fig. 5-9 Sigma-delta modulation used in an A/D converter

As we know, most music signals have a large dynamic range. To ensure adequate rendering over the whole range, which is, for example, equivalent to 16 bit PCM, for a one bit sigma-delta modulator, either very high order of loop filter or very high oversampling ratio should be used. In many cases, the low music volumes predominate. Fig. 5-10 shows an example of the magnitude distribution from a 15-second piece of music. The full scale value of magnitude is 32768. It shows that

large magnitude values are relatively rare. It indicates that over most of time we will obtain very poor signal-to-noise ratio. The average signal-to-noise ratio of the system to music signals will be very low. If we reduce the noise for the predominant weak signals, even at the expense of an increase in noise for the rarely occurring strong signals, the average signal-to-noise ratio can be improved. This goal can be gained by using an A/D converter which consists of an adaptive sigma-delta modulator.

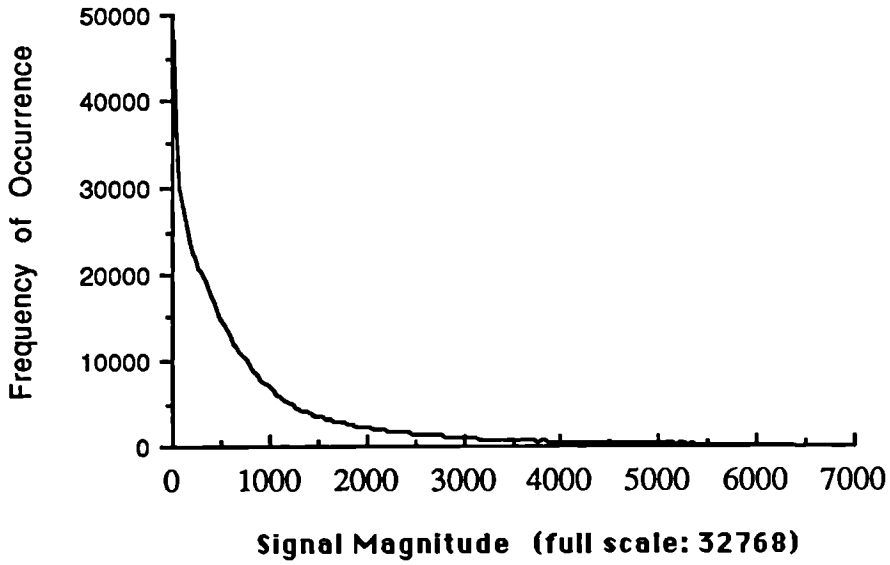



Fig. 5-10 Magnitude distribution of a 15-second piece of music

It has already been shown that at the lower level of input, the adaptive sigma-delta modulator is equivalent to higher bit PCM. It means that the quality for small signals can be improved at the same oversampling ratio, or can be made the same as that of a higher oversampling ratio. Fig. 5-11 shows the results of comparison between 128 oversampling ratio, fixed 3rd order and 64 oversampling ratio,adaptive 3rd order sigma-delta modulators. Both of the results are those which have been converted into 16 bit PCM. From Fig. 5-11, it can be seen that when input is below -15 dB, both of

them have nearly the same SNR. The oversampling ratio can be reduced by half with an adaptive quantiser by sacrificing some SNR of large signal while keeping the same quality when input level is lower.

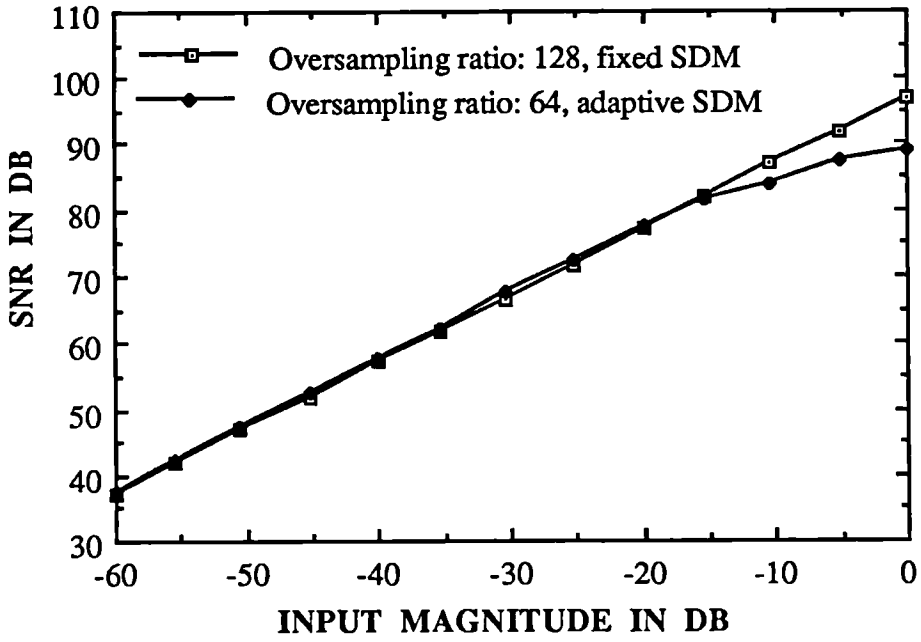



Fig. 5-11 Comparison between 3rd order, 128 oversampling ratio, fixed and 64 oversampling ratio, adaptive sigma-delta modulators ( converted into 16 bit PCM ) ( input: 10087Hz sinewave; oversampling ratio: 64; Nyquist sampling frequency: 44.1 kHz; 1-bit quantiser. )

From the system order point of view, Fig. 5-12 gives the results of comparison between fixed 3rd order and adaptive 2nd order SDMs with the same oversampling ratio of 128. It shows that the order of a SDM system can be reduced by using an adaptive quantiser while maintaining the same quality for small signals, which also leads to better behaviour in system stability.

The above results indicate that the overall signal-to-noise ratio can be improved by reducing the noise for the predominant weak signals, at the expense of an increase in noise for the rarely occurring strong signals.

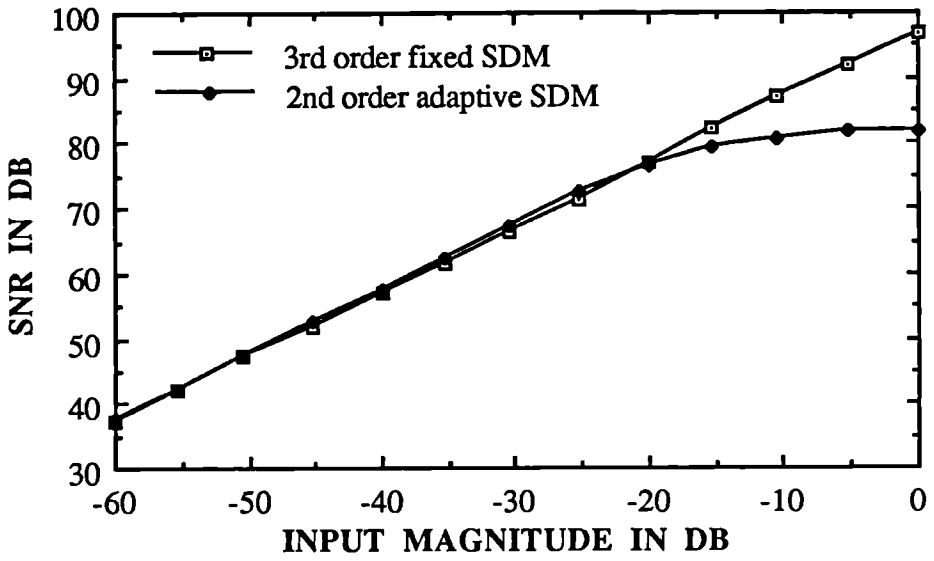



Fig. 5-12 Comparison between 128 oversampling ratio, 3rd order fixed and 2nd order adaptive sigma-delta modulators ( converted into 16 bit PCM )

Adaptive SDM can also be used in a similar way as a D/A converter. Because the basic principle is the same, the results similar to Fig. 5-11 and 5-12 should be possible to be obtained.

## 5.6 Music signal tests

Music signals can be quite different from sinusoidal signals. They usually have a large dynamic range in magnitude and vary dramatically with time. A waveform representing a typical music signal is shown in Fig. 5-13. It is evident from this figure that the properties of the music signal change with time. For example, there is significant variation in the peak amplitude of the signal, and there is considerable variation of fundamental frequency. In this section we describe the tests we have carried out on music signals for adaptive sigma-delta modulation (ASDM).

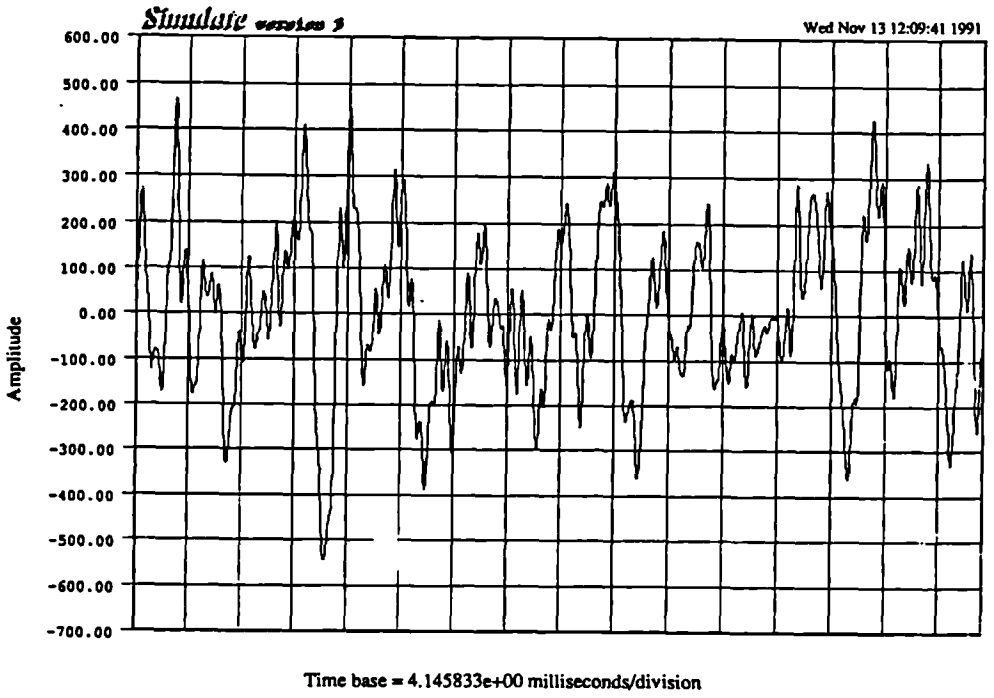



Fig. 5-13 Waveform of a piece of music

### 5.6.1 Testing procedures

Since the oversampled analogue music signal is not directly available in the computer and also A/D interface which has more than 16 bit resolution is not available, ASDM as an exact A/D converter cannot be tested. In order to test the function of the system, the following procedures have been used. A 15-second piece of music is inputed through 16-bit A/D interface to the computer, at sampling rate 48 kHz. The sequence with 16-bit resolution is then interpolated by a factor of 64 and used as an input to the ASDM system. Afterwards, the output signal of the ASDM is lowpass-filtered and decimated down to the 48 kHz. At this stage the frequency and time domain analysis and other measurements by computer simulations can be carried out. Finally, the decimated signal can be converted back to 16-bit digital signal. The results can be judged by listening through playing-back system. The whole simulation diagram is shown in Fig. 5-14. The system inside the dashed frame actually functions as a D/A converter. The whole simulation system, in fact, contains two A/D and two D/A converters.

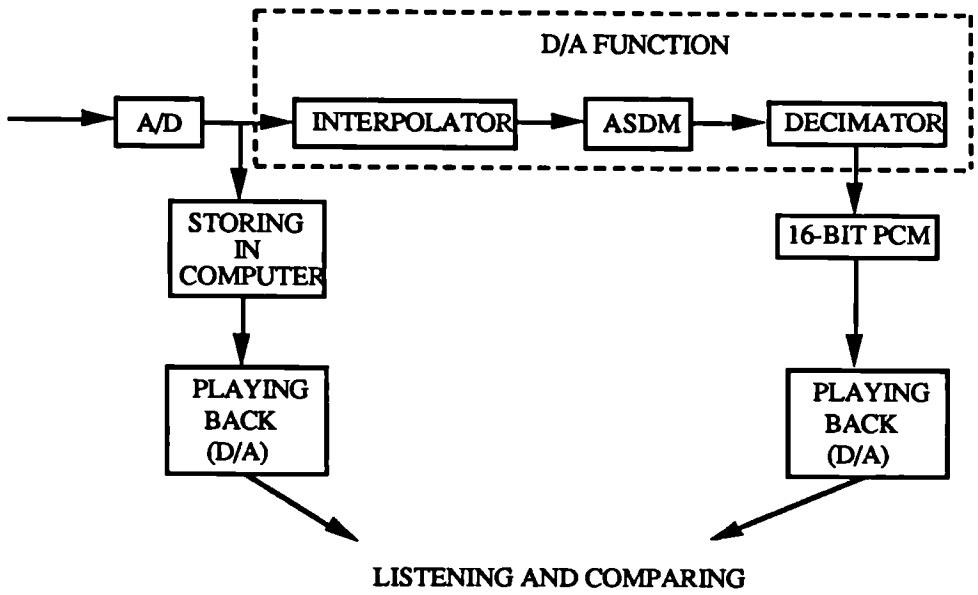



Fig. 5-14 Diagram for music signal test

## 5.6.2 Short-time Fourier analysis

As mentioned above, the properties of the music signal vary with time. However, we can assume that the properties are relatively constant over a short period of time. This assumption leads to a variety of "short-time" analysis methods [29]. The most useful for the music signal test of ASDM is short-time Fourier analysis.

In order to study the spectral properties of a music signal, it is convenient to introduce the concept of a time-varying Fourier transform. Supposing that x is the music signal, a useful definition of its time dependent Fourier transform is

$$X_k(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(k-m)\, x(m)\, e^{-j\omega m} = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})\, e^{-j\theta k}\, X(e^{j(\omega-\theta)})\, d\theta \qquad (5\text{-}5)$$

where w(k-m) is a real "window" sequence which determines the portion of the input signal that receives emphasis at a particular time index, k. Strictly speaking the normal Fourier transform of a music signal does not exist. However, equation (5-5) is meaningful if we assume that $X(e^{j\theta})$ stands for the Fourier transform of a signal whose basic properties either continue outside the window or which is zero outside the window. Thus the time dependent Fourier transform can be interpreted as a smoothed version of the Fourier transform of the part of the signal within the window.

The time dependent Fourier transform is clearly a function of two variables: the time index k, and the frequency variable $\omega$. For fixed k, $X_k(e^{j\omega})$ has the same properties as a normal Fourier transform. The sufficient condition for the existence of the time-dependent Fourier transform is that the sequence x(m)w(k-m) is absolutely summable for all values of k. If, as is often the case, w(k-m) is of finite duration, then this condition is clearly satisfied.

The shape of the window sequence has an important effect on the nature of the time-dependent Fourier transform. It is clear from (5-5) that for faithful reproduction of the properties of X ($e^{j\omega}$) in $X_k(e^{j\omega})$, the function W ($e^{j\theta}$) should appear as an impulse with respect to X ($e^{j\omega}$). For example, a rectangular window has relatively narrow main lobe width compared with triangular, Hanning, Hamming, Blackman windows etc., but it has large side lobes which produce a "ragged" or noisy spectrum and tend to offset the benefits of the narrower main lobe. As a result, such windows are rarely used in speech or music signal spectrum analysis.

In evaluating the frequency domain properties, a "4-term" window described by Nuttall [47] is chosen:

$$w(k) = 0.338946 + 0.481973\cos(\pi k/N) + 0.161054\cos(2\pi k/N)$$

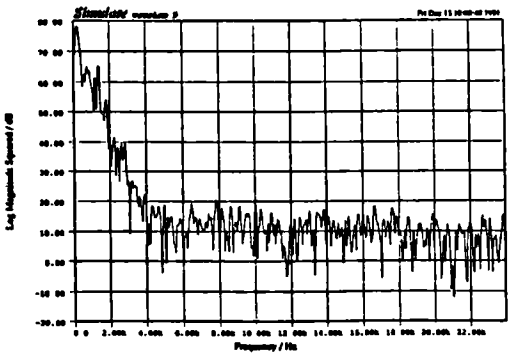
$$+ 0.018027\cos(3\pi k/N), \qquad k = -N, ..., -1, 0, 1, ..., N-1$$

where N is the half length of the window. This window is shown to have very low maximum side lobes of -82.6 dB and a fast side lobe decay at a rate of 30 dB/octave.

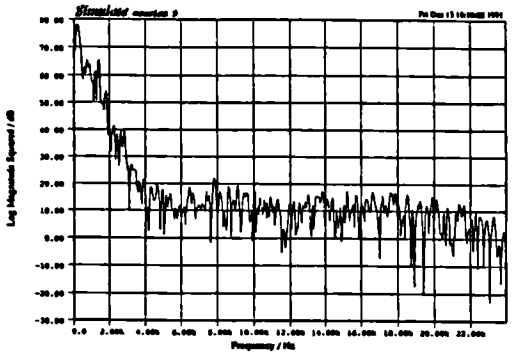
Fig. 5-15 shows a set of spectra comparison between the original and reconstructed signal by using the 3rd order ASDM with 64 oversampling ratio over two different period of time. There is no big difference between the original and the reconstructed except for the magnitude droop at high frequencies for the reconstructed signal. This is probably because of the nonideality of the low-pass filter in the sigma-delta demodulator.

Fig. 5-16 gives the examples of spectral comparison between fixed and adaptive SDMs over about 10.7 ms period of time. The maximum magnitude over this period is about 36 dB down from the full scale. The spectra clearly show that the small signal

benefits very much from the adaptive sigma-delta modulation. Also, it is more advantageous when using lower order systems.

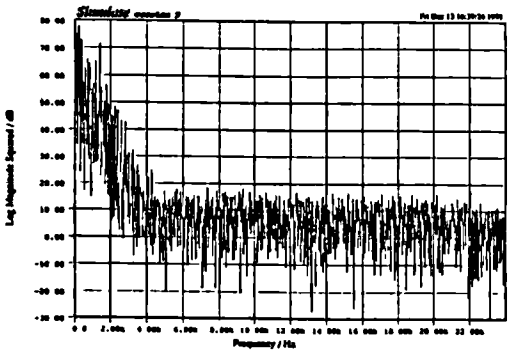


(a) Original

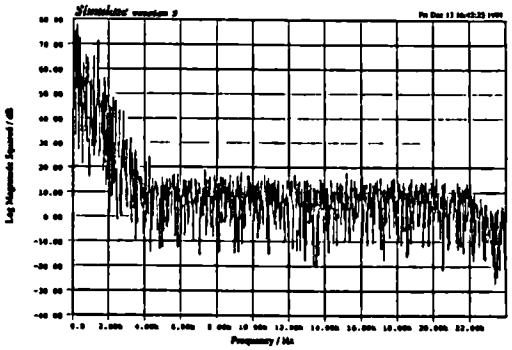


(b) Reconstructed

Magnitude spectra over 1024 samples (app. 21.3 ms)



(c) Original

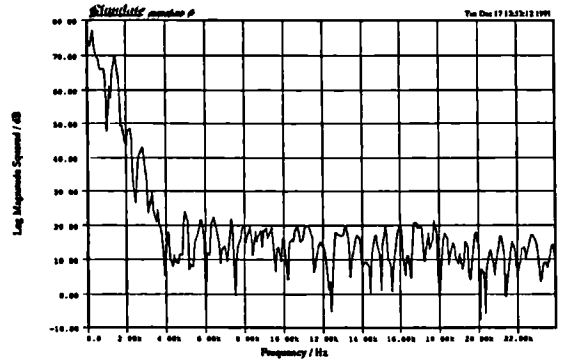


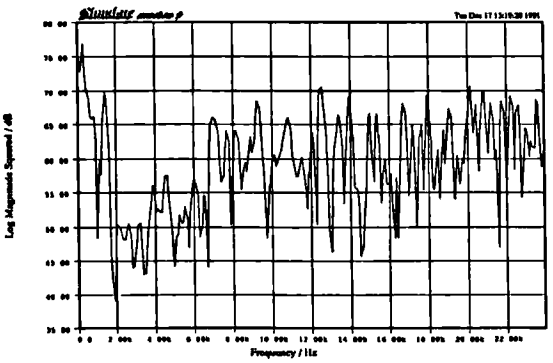
(d) Reconstructed

Magnitude spectra over 8192 samples (app. 170 ms)

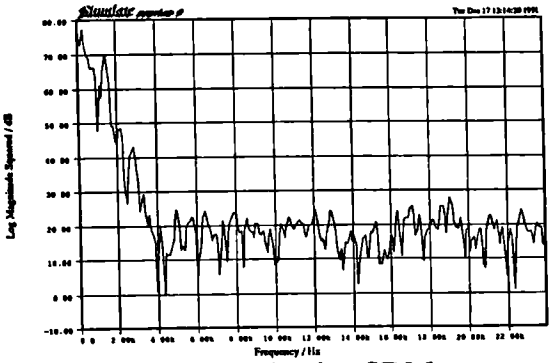
Fig. 5-15 Comparison of spectra between the original and reconstructed music signals

(a) original

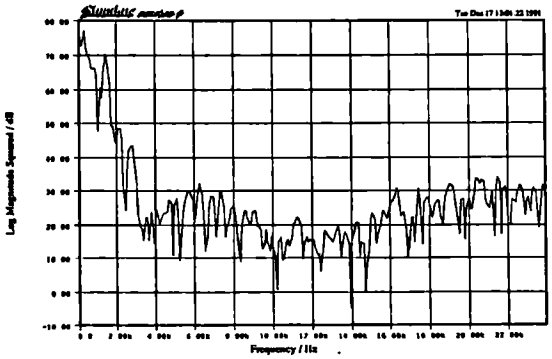




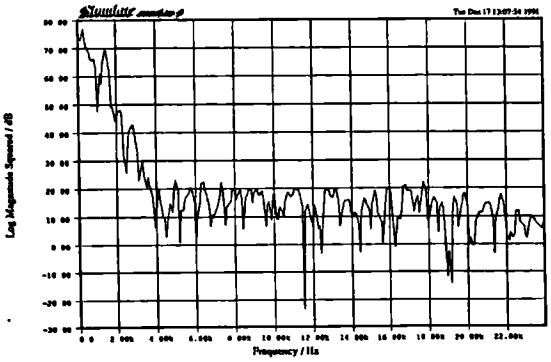
(b) 1st order fixed SDM



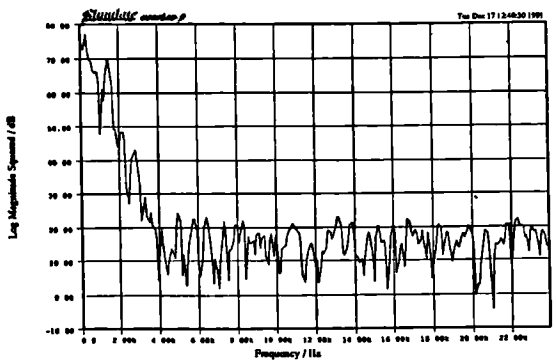
(c) 1st order adaptive SDM



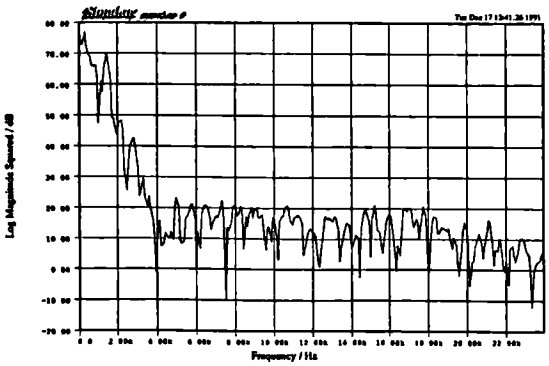
(d) 2nd order fixed SDM



(e) 2nd order adaptive SDM



(f) 3rd order fixed SDM

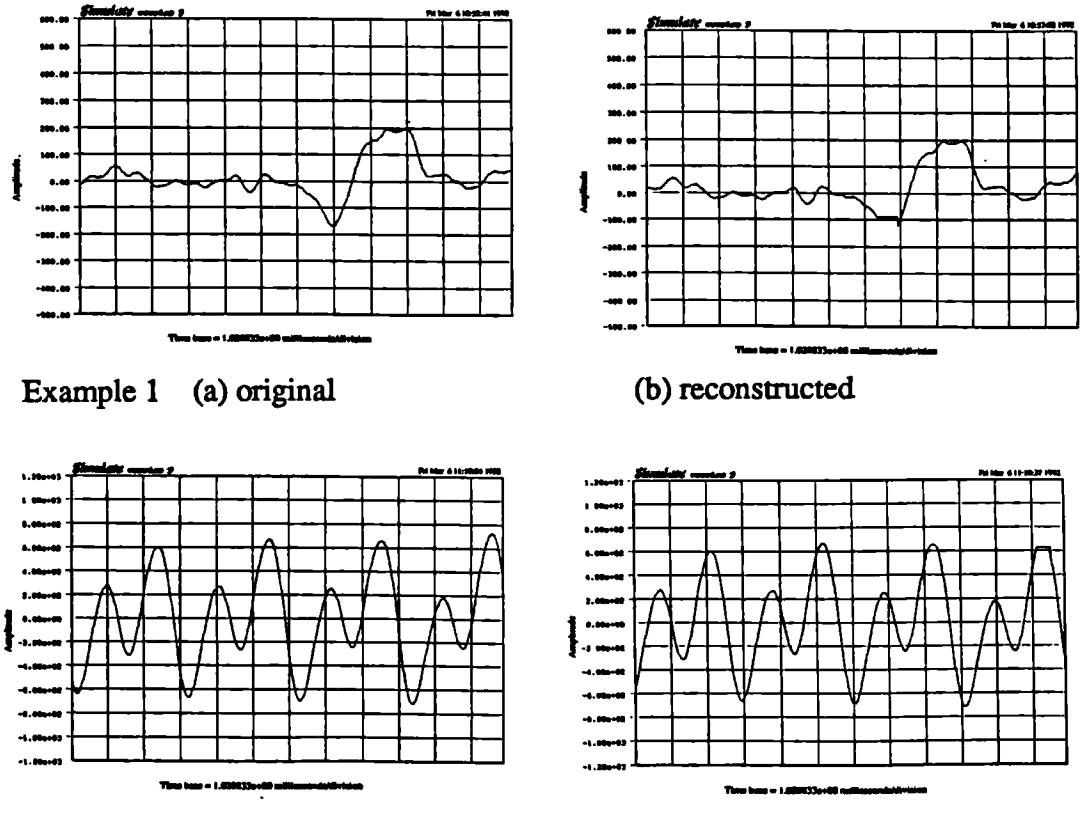


(g) 3rd order adaptive SDM

Fig.5-16 Magnitude spectra over 512 samples

## 5.7 Effect of adaptation speed on music signals

The adaptation logic mentioned in Section 5.3 can operate very well when input is a sinewave signal because the peak magnitude of the signal will be constant. For a music signal, the magnitude usually changes dramatically with time so that if the adaptation is calculated in the current block and used for the coming block, and if the maximum magnitude of the coming block is greater than the current, severe distortion will occur. Fig.5-17 gives two examples. The major distortion caused by this problem is overload distortion which is much worse than the noise caused by coarse quantisation.



Example 1 (a) original (b) reconstructed

Example 2 (a) original (b) reconstructed

Fig. 5-17 Illustration of the effect of adaptation speed

One easy way of solving this problem is to set the quantisation level larger than it should be. In order to avoid the overload distortion completely, the quantisation level should be set large enough. This will introduce more quantisation noise. Computer simulations show that for the particular music example, the adapted quantisation level has to be three times larger than the optimal level described in Chapter 3 for preventing overload so that the signal-to-noise ratio will decrease by about 9.5 dB.

Another way seems that we can apply the logic of calculation from current block of data to the same block. However, as is shown in Fig. 5-18, this method will introduce delay. Furthermore, it needs two sets of analogue loop filter and quantiser, which increases the complexity of the circuits.

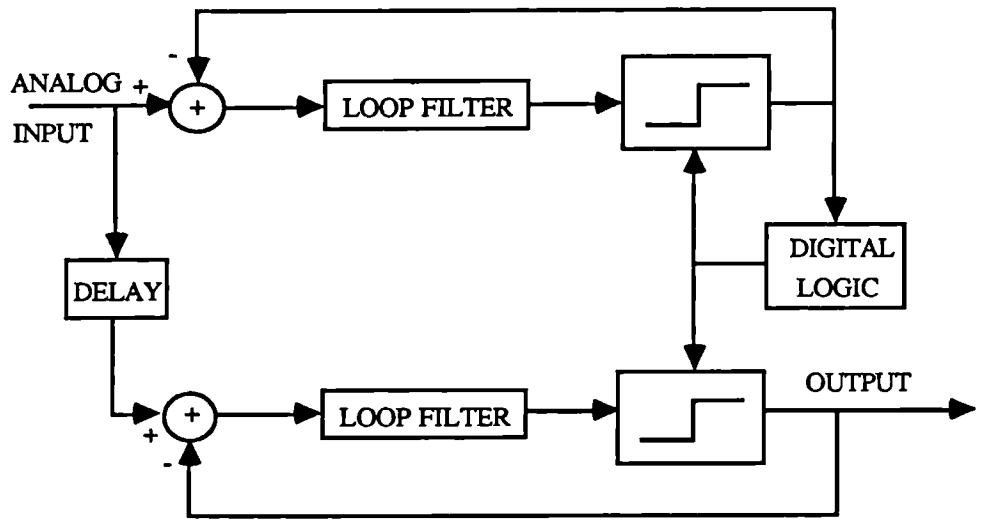



Fig. 5-18 A possible way of avoiding mis-tracking

A practical way is to change the adaptation speed. One of the major factor of adaptation speed is the block (or buffer) length K. When K=1, it is called

instantaneous adaptation. when K is approximately equal to the period of the signal (if the signal is periodic or pseudo-periodic), it is called syllabic adaptation.

Generally speaking, instantaneous adaptation system has the optimum quantisation level selected anew for each sample. It would appear to be optimum. However, in the particular sigma-delta modulation system, the backward adaptation logic is used and the magnitude is detected after the decimator. The decimator will introduce some delay so that the sample-based adaptation logic will be applied after a delay of several samples, therefore it is very inaccurate. Furthermore, because the quantisation level changes with the signal level, the system might produce modulation noise. This can be extremely disturbing, especially when the signal is low frequency and high amplitude. The quantisation level will change many times within one period. The quantisation noise, being wide band, cannot be masked by the signal. For example, there can be 40 dB of quantisation level change, and hence 40 dB of quantisation noise modulation. The quantisation noise will follow an almost inaudible signal. In other words, the signal can hardly be detected but the noise might be very disturbing.

The syllabic system is better than the instantaneous with low-frequency signals because the quantisation level is held constant within one period and hence the quantisation noise is constant. The problem is how long the block length should be. On the one side, it is assumed that over a short period of time, the music signal is stationary. This suggests that we have to adapt the quantisation level quickly in order to follow the change of the signal, especially, when the signal changes from low magnitude to high magnitude, as is shown in Fig. 5-17. Otherwise, severe overload distortion will occur. However, the block size cannot be too small. Computer simulations show that severe error will happen in the situation shown in Fig. 5-19. If

we calculate the maximum value over Block i and apply it to Block i+1, the severe overload distortion will occur. In fact, in this example, the stationary period of the signal is longer than the block size. The signal changes from large to small and back to large magnitude again. Therefore, we wish that the quantisation level decays with a certain delay. This indicates that the block length should be sufficiently long.

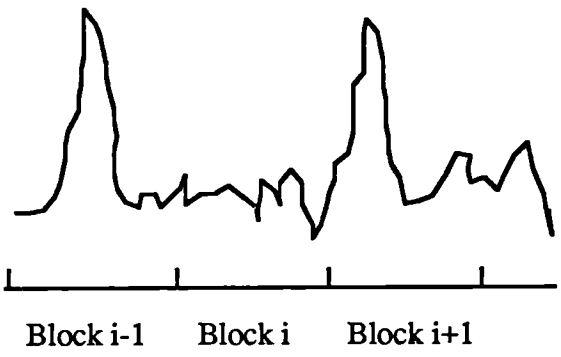



Fig. 5-19 Severe overload distortion will occur if the maximum value over Block i is calculated and applied to Block i+1

To keep the information over a longer period of time and still to adapt the quantisation level over a relatively short time, we can carry out the calculation over the longer one but the adaptation over the short one, as is shown in Fig. 5-20. The calculation block will shift along the time axis. In this way, if a large magnitude occurs, the quantisation level will follow it more quickly by each time shifting the calculation block a short period of time (the length of the adaptation block). Also, it will keep the large quantisation level over a long calculation block to prevent the situation in Fig.5-19. Therefore, we establish a fast attack time but a slow release time. To reduce the delay effect caused by the decimator, a very simple comb filter is used. The functions are simply time average and decimation, as shown in equations (2-34) and (2-35). It is a FIR filter whose delay is N/2, where N is the oversampling ratio. The delay is only half a sample at Nyquist rate and can be considered very small. In order

to have a safety margin, we also set the quantisation level slightly larger than it should be according to the adaptation logic. By using these two ways together, the distortions shown in Fig. 5-17 for this particular piece of music disappear.

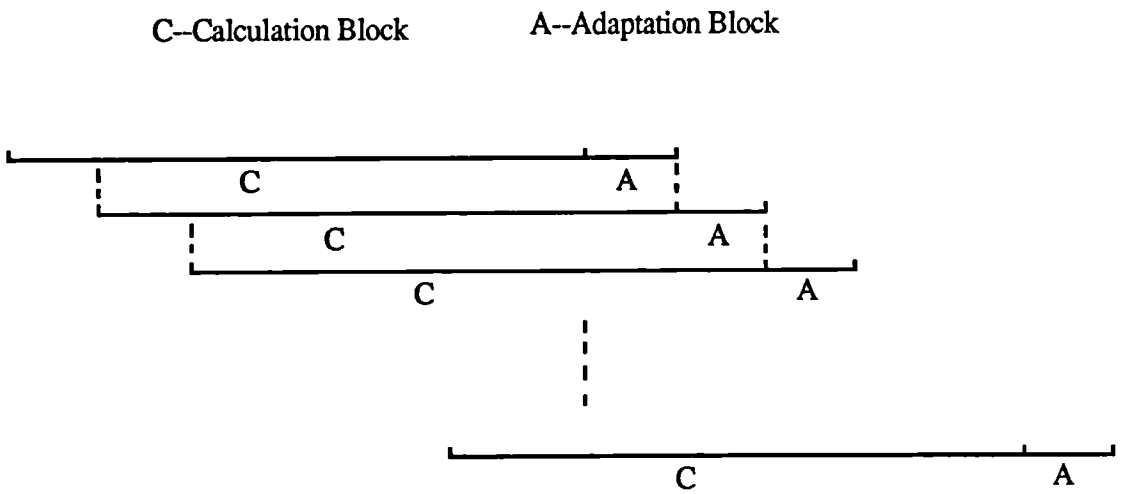



Fig. 5-20 A longer calculation but a shorter adaptation block

## 5.8 Quantised adaptation levels

One possible way of implementing the adaptation is by using a multiplying D/A converter (MDAC) in the feedback path as shown in Fig. 5-21. The output of MDAC will be the result of multiplication of analogue reference voltage ( the output of the quantiser ) and the output of the digital logic block.

The structure in Fig. 5-21 includes a multi-bit D/A converter which reduces the advantages of VLSI implementation of single bit SDM. However, it is still different from multi-bit SDM which is shown in Fig. 5-22. It contains not only a multi-bit D/A but also a multi-bit A/D converter. As we know, a multi-bit A/D conversion process is generally more complex and time consuming than a D/A process. Several important

types of A/D converters (ADCs) such as successive-approximation, digital-ramp ADC etc., utilise a D/A converter as part of their circuitry [48]. Therefore, an adaptive SDM seems still simpler than a multi-bit fixed SDM.

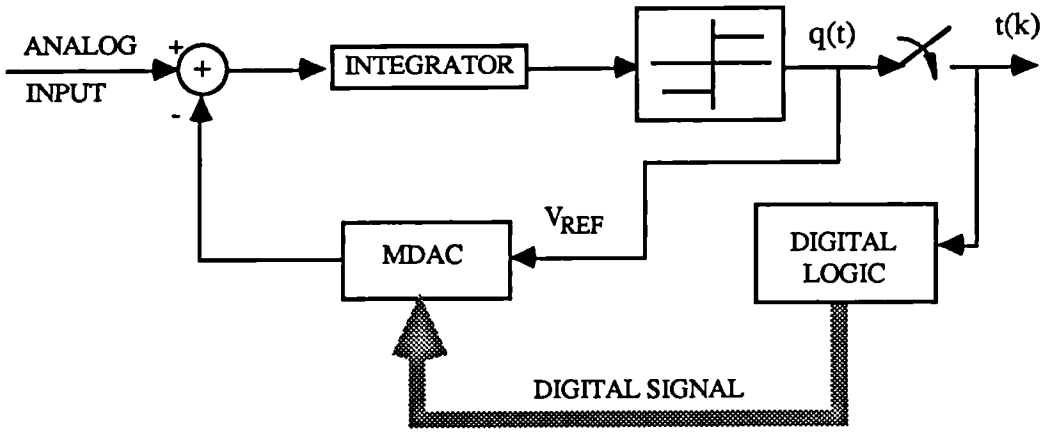



Fig. 5-21 Using MDAC to implement the adaptation logic

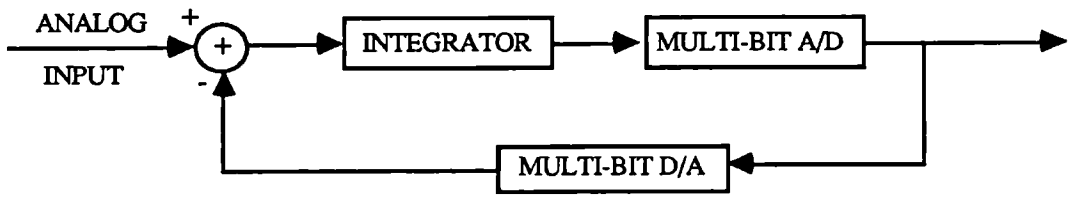



Fig. 5-22 Block diagram of a multi-bit SDM

In the previous sections we have discussed the ideal case of adaptation, that is, the quantisation level can be changed continuously. All the computer simulations have been carried out based on this continuous model. However, the "digital logic" block

and the MDAC in Fig. 5-21 will be of limited length. It indicates that the quantisation level can only be adapted within the certain limited discrete levels. The quantisation on adaptation level: d=cE (see Section 5.3) can be considered equivalent to the quantisation effect on the amplitude of the input signal: E, because c is a constant. Supposing that K-bit uniform quantisation is carried out on E within the range ($E_{max}$, 0), the quantisation logic is as follows. If

$$i\frac{E_{max}}{2^K} < E \le (i+1)\frac{E_{max}}{2^K}, \qquad i = 0, 1, ..., 2^K-1$$

the quantised E is

$$E_{quan} = (i+1)\frac{E_{max}}{2^K}, \qquad i = 0, 1, ..., 2^K-1$$

Fig. 5-23 gives the quantiser characteristic when K equals 2. The upper bound is chosen for quantisation level instead of the middle value between the upper bound and the lower bound as is in the usual midrise quantiser. The purpose for this is to avoid the overload distortion happening in the whole SDM system. If midrise or midtread quantiser is chosen, $E_{quan}$ will be sometimes smaller than E so that $d=cE_{quan}$ is smaller than $d_{opt}$ in Fig. 5-3, which is the case of overload.

Suppose that after quantisation on adaptation level, the equivalent error on E is $\Delta E$ so that the real quantisation level in this case will be

$$d_{real} = c\,(E + \Delta E), \qquad 0 \le \Delta E < E_{max}/2^K \tag{5-6}$$

Replacing d in (5-3) with (5-6), the real signal-to-noise ratio ($SNR_{real}$) will be

$$SNR_{real} = \frac{3E^2}{2c^2(E+\Delta E)^2} SNR_{enhancement}$$

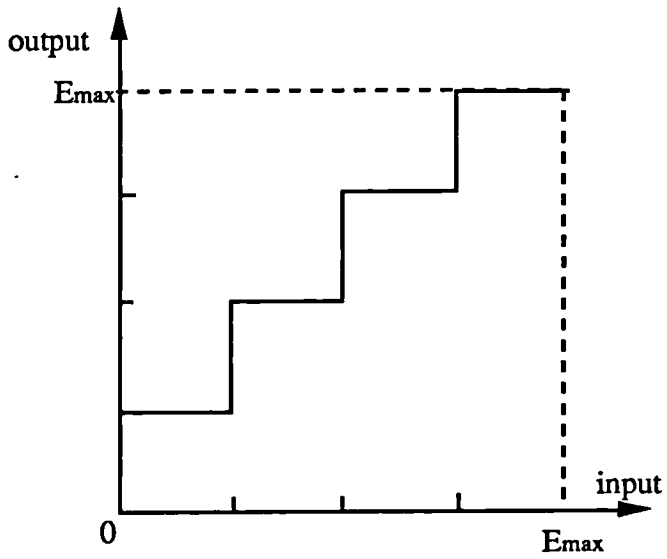



Fig. 5-23 Quantiser characteristic when K=2

Considering (5-4) as the ideal case $SNR_{ideal}$, $SNR_{real}$ can be expressed as

$$SNR_{real} = SNR_{ideal} \frac{1}{(1+\Delta E/E)^2}$$

The real SNR in dB will be

$$SNR_{real}(dB) = SNR_{ideal}(dB) - 20\log(1+\Delta E/E)$$

In the case of $\Delta E=0$, $SNR_{real}$ will be the same as $SNR_{ideal}$. We define $\Delta SNR(dB)$ as

$$\Delta SNR(dB) = 20\log(1+\Delta E/E) < 20\log[1+E_{max}/(E2^K)] \qquad (5\text{-}7)$$

It indicates that the bigger the K, the smaller the upper bound of $\Delta SNR(dB)$. When K approaches infinity, the difference $\Delta SNR$ becomes zero. It also shows that the smaller the amplitude E, the bigger the upper bound. Fig. 5-24 gives several upper bound curves of $\Delta SNR$ versus input magnitude $E/E_{max}$ (dB) for different bit number K.

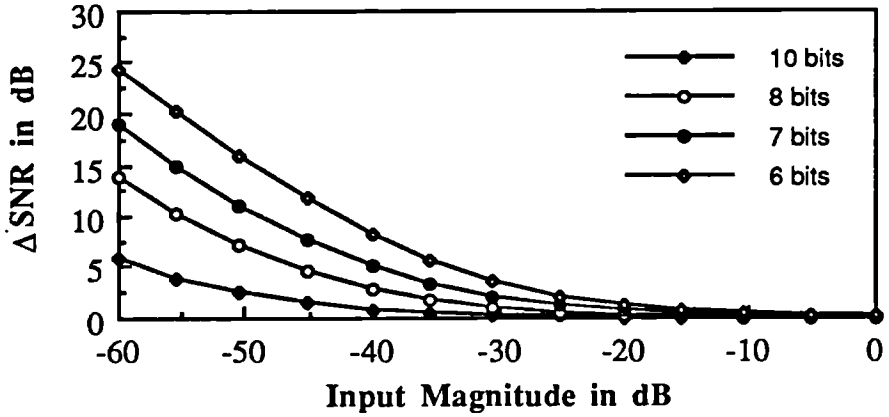



Fig. 5-24 Upper bounds of the loss in SNR versus input magnitude $E/E_{max}$

Fig. 5-25 gives the simulation results of the third order adaptive SDM (ASDM) with quantised adaptation levels. Compared with the ASDM with continuous adaptation level, the curves of signal-to-noise ratio versus input level start to drop after the input level falls to some certain levels. The input level at which the SNR starts to drop becomes higher as the length (bit number) of the D/A part in the loop decreases.

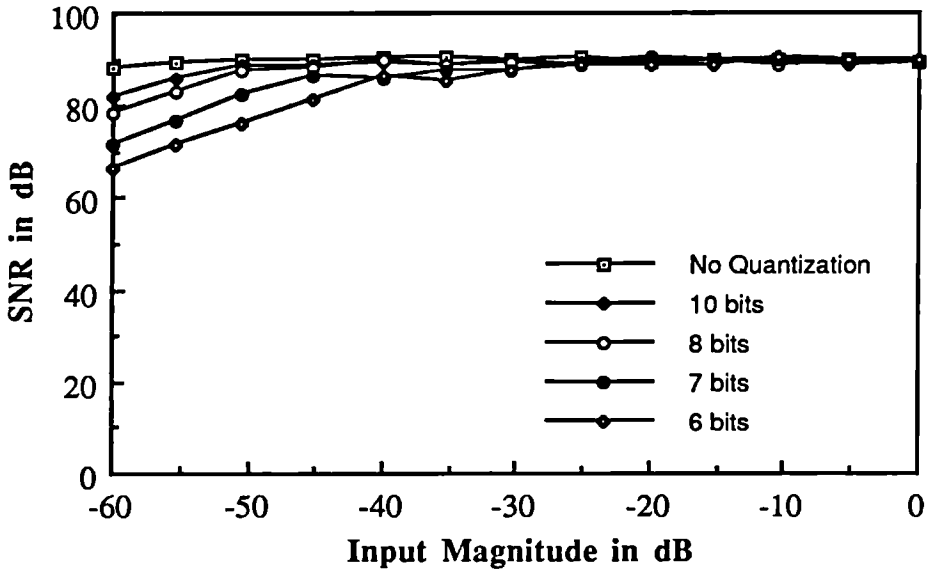
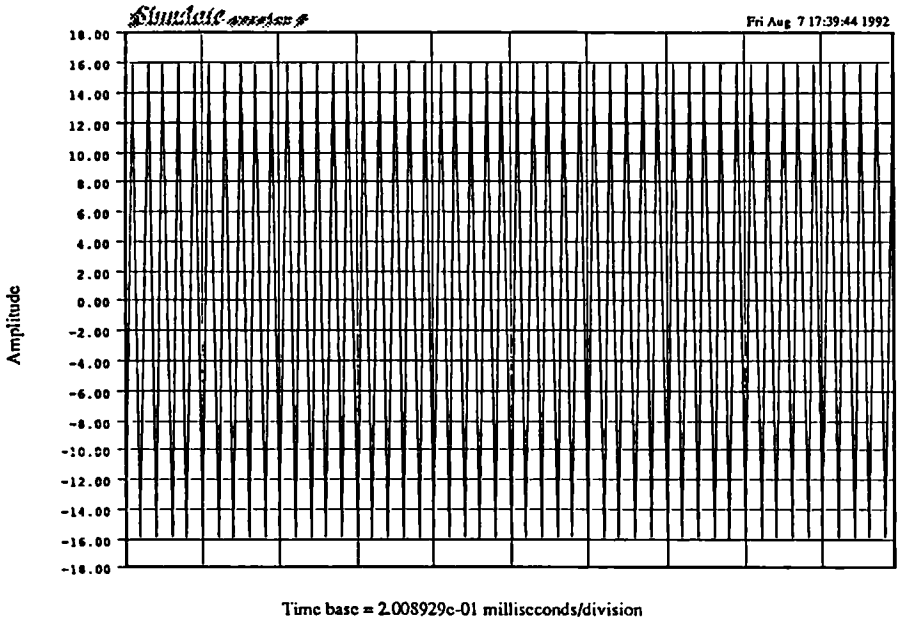



Fig. 5-25 SNR curves of the 3rd order adaptive SDM with quantised adaptation levels

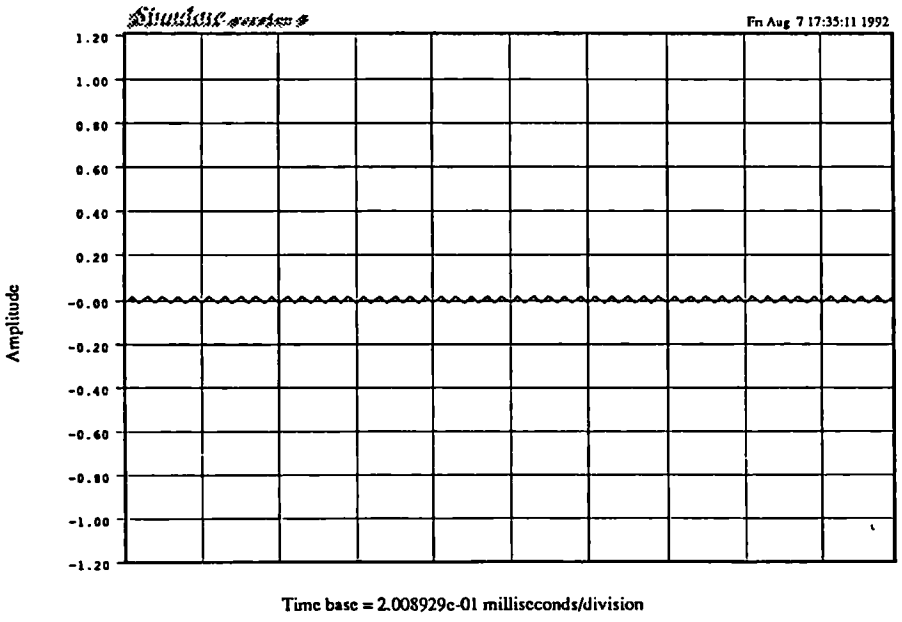
## 5.9 Simulations of idle channel noise

As is shown in Section 3.6 of Chapter 3, the upper bound of the idle channel noise is proportional to the quantisation level. In the adaptive SDM, the quantisation level varies in the range $[d_{min}, d_{max}]$. $d_{max}$ can be determined based on the system limitation or the possible maximum input magnitude. In a SDM with a fixed quantiser, in order to avoid the overload distortion, $d = d_{max}$ is chosen and kept constant all the time. Therefore, according to equation (3-11), the upper bound of the idle channel noise for the fixed SDM will be $2d_{max}/N$ for the system in Fig. 3-8. If an adaptive quantiser is chosen, the upper bound of the idle channel noise will reduce to $2d_{min}/N$.

If $d_{max}$=1000.0 and $d_{min}$=1.0, the idle channel noise will be reduced by a factor of about 1000. Fig. 5-26 (a) and (b) show the time-domain waveforms of the idle channel noise of the first order fixed and adaptive SDM respectively, where the filter coefficients are $b_1$=1.0, $a_1$=0.0 and the oversampling ratio is 63. Fig. 5-27 gives the comparison of the two waveforms when $b_1$=1.0 and $a_1$=-0.0059375.



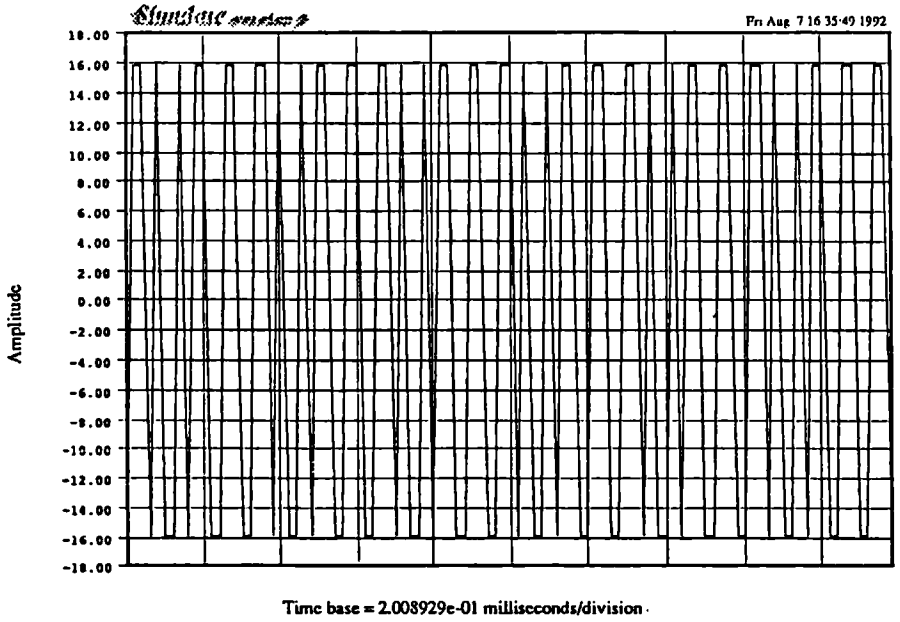

(a) The fixed SDM



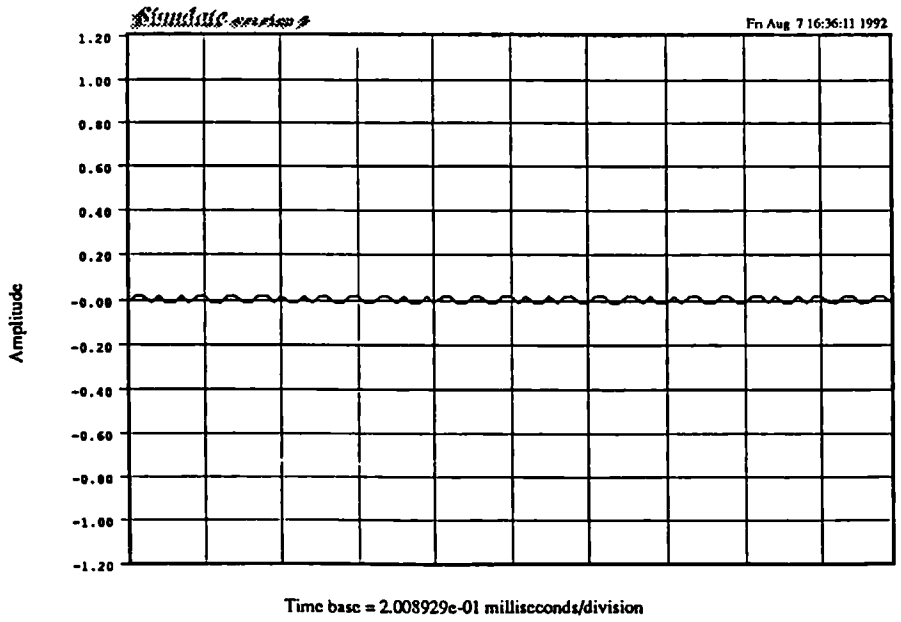

(b) The adaptive SDM

Fig. 5-26 Time-domain waveforms of the idle channel noise

when $b_1$=1.0 and $a_1$=0.0




(a) The fixed SDM




(b) The adaptive SDM

Fig. 5-27 Time-domain waveforms of the idle channel noise

when $b_1$=1.0 and $a_1$=-0.0059375

## 5.10 Conclusions

In this chapter, we have discussed SDM systems with the adaptive quantiser. The main points and the conclusions are summarised as follows:

1. The feedback adaptation has to be chosen mainly because the input will be analogue if a SDM is used as an ADC.

2. It is impossible to estimate the magnitude of the input of the one-bit quantiser from its output. To solve this problem, the output of the equivalent quantiser is used to estimate its own input magnitude, i.e., the input magnitude of the SDM system.

3. The computer simulations show that an adaptive SDM gives much wider dynamic range than a fixed one.

4. By defining the minimum quantisation level $d_{min}$ appropriately, the idle channel noise can be reduced to a very low level.

5. When a SDM with adaptive quantiser is used as an ADC, and the output is finally converted to linear PCM, the oversampling ratio, or the order of the loop filter can be reduced while maintaining the same SNR for small signals at the expense of an increase in noise for the rarely occurring strong signals.

6. In order to tailor the adaptive SDM to music signals, a fast attack time, but a slow release time, are established by using a long calculation block but a short adaptation block.

7. The proposed design procedure of adaptation also applies to the noise shaper structure described in Section 2.4, because it is in essence the same device as the SDM. The results from the adaptive noise shaper are expected to be similar to the corresponding results in this chapter.

# 6

# ADAPTIVE LOOP FILTER FOR SIGMA-DELTA MODULATION

## 6.1 Introduction

In Chapter 3, it has been shown by both theoretical analysis and computer simulations that the possible maximum amplitude without causing overload distortion is always smaller than the quantisation level for sinusoidal input. In other words, the input must be lower than a certain level in order to obtain a reasonable signal-to-noise ratio (SNR). This is similar to the case of PCM systems for preventing overload distortion. As is also mentioned in Chapter 3, the maximum possible input level decreases with the increase of the order of loop filter. If the quantisation level is set to be one (normalised value), the input magnitude should be less than 0.6 for the 2nd order sigma-delta modulator (SDM), and less than 0.29 for the 3rd order SDM.

If when the input level increases the loop filter can be adapted so as to prevent sudden drop of the SNR, then the dynamic range can be increased. Because of the difficulty of maintaining the SDM system stability, no one has tried to work on the

SDM with adaptive filter. At least, no publications on it have been seen so far, to the author's knowledge. This chapter is an attempt to adapt some of the coefficients according to the input magnitude so as to improve the dynamic range of the system.

## 6.2 Adaptation logic

Generally speaking, an adaptive filter has an adaptation algorithm which enables the transfer function to track, in a useful manner, some feature of its external environment. Specifically, the adaptation algorithm monitors the external influence or environment of the filter and controls its transfer function by varying its parameters.

The type of adaptation algorithms usually can be classified as block or instantaneous. In the block algorithms, the input signal is divided in time into blocks, and each block is processed independently (although sometimes there is commonly some overlap between the adjacent blocks). Within each block, the optimum parameters of the adaptive filter can be determined. The result is a new set of filter parameters at the end of each block. In the sample-based, i.e., the instantaneous algorithms, on the other hand, the adaptive algorithm is implemented as a continuous operation, so that a new set of parameters is generated at each input data sample. In a similar way to adaptive quantiser in Chapter 5, the adaptation method for the loop filter can also be divided into two different types: feed-forward and feedback ones.

The general idea of adaptive loop filter is that the coefficients of the loop filter change according to the magnitude of the input. Suppose that the whole SDM is a digital system or it is used as a DAC so that the input is a digital signal and the loop

filter is a digital filter. Otherwise, if SDM is used as an ADC, it is difficult to adapt the loop filter because it is an analogue one. Because of the digital input, we may choose the feed-forward adaptation logic so that more accurate estimation can be obtained. In particular, if SDM is used as a DAC, then the output of it is analogue so that it is more difficult to obtain the estimation by using feedback adaptation. Therefore, the feed-forward adaptation is chosen. Fig. 6-1 shows the block diagram of a sigma-delta modulator with an adaptive loop filter. If the system contains an interpolator, the adaptation logic can be carried out at lower sampling rate, which is shown in Fig. 6-2. Therefore, the coefficients of the loop filter are time varying.

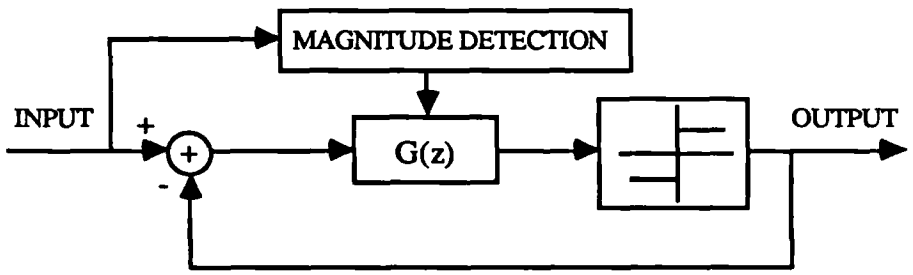



Fig. 6-1 Block diagram of a sigma-delta modulator with an adaptive loop filter

The main idea is to attempt to extend the dynamic range of the high order sigma-delta modulators by adapting the coefficients of the loop filter. It is important to operate the sigma-delta modulation system with the appropriate coefficients of the loop filter so that the system will be stable. This indicates that the coefficients cannot be arbitrarily adapted.

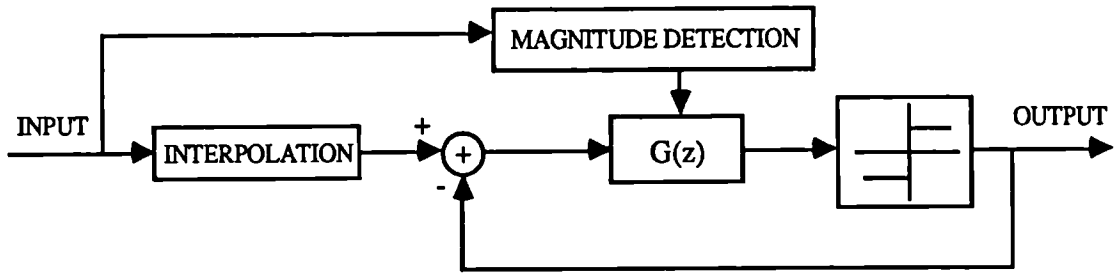



Fig. 6-2 Block diagram of a sigma-delta modulator with an adaptive loop filter working at lower sampling rate

Computer simulations show that $\{b_i\}$ coefficients mainly affect the stability and signal-to-noise ratio while $\{a_i\}$ coefficients do not play key roles in a SDM system. As a result, $\{a_i\}$ coefficients will be constants during the adaptation process; only $\{b_i\}$ coefficients are considered to be adapted. The corresponding block diagram is depicted in Fig. 6-3. Based on the optimal coefficients of high order loop filter, for example, the 4th order loop filter, which have been discussed in Chapter 3, and starting from the maximum possible input magnitude of sinewave signal for the particular high order system, we increase the input magnitude gradually and search the optimal $\{b_i\}$ coefficients under the increased level by using the optimisation method described in Chapter 3 and Appendix B.

Supposing that the maximum input magnitude of sinewave is one for the fixed optimum 4th order SDM, the optimisation simulations are carried out for different input magnitudes. The magnitude values are divided into six different ranges and the simulation results of optimal $\{b_i\}$ for those six ranges are listed in Table 6-1. Adaptation logic can be implemented by detecting the magnitude of the input and then updating the filter coefficients by a table-look up method.

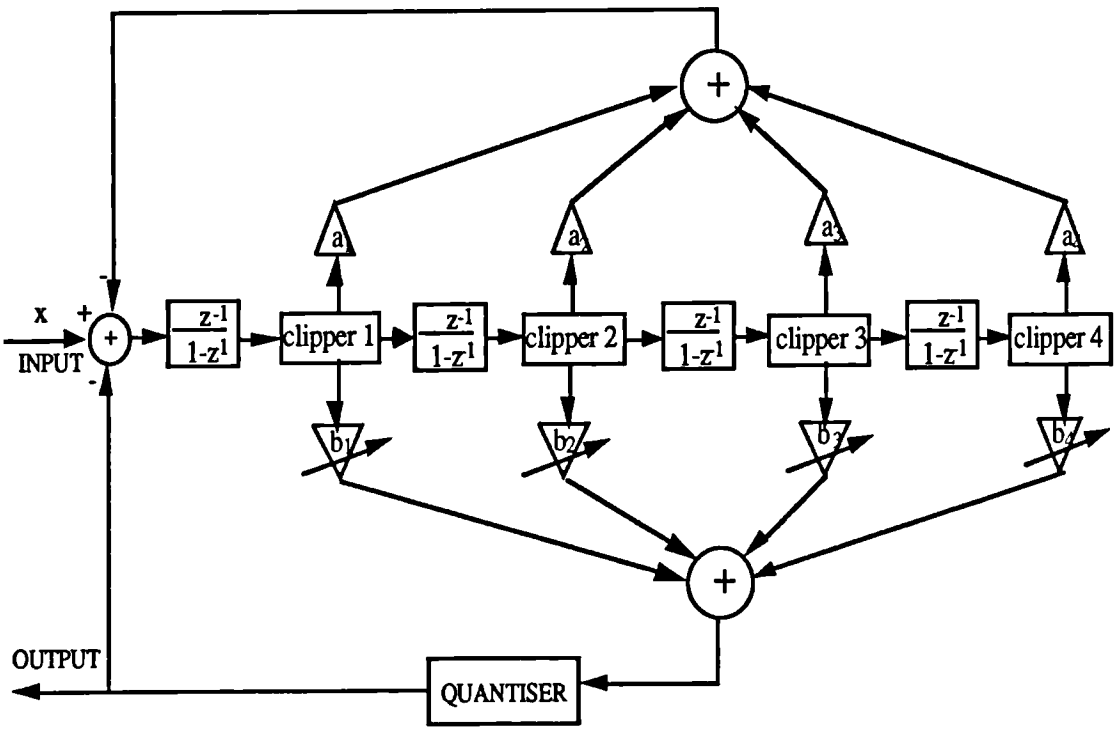



Fig. 6-3 A SDM with a fourth order adaptive filter

Table 6-1 Simulation results of optimal $\{b_i\}$ for different ranges of input magnitude, where $b_1$=1.0

| Input Magnitude A | $b_2$ | $b_3$ | $b_4$ |
|---|---|---|---|
| $A \leq 1.0$ | 0.4865 | 0.110896 | 0.020034 |
| $1.0 < A \leq 1.167$ | 0.467872 | 0.085908 | 0.017534 |
| $1.167 < A \leq 1.333$ | 0.405257 | 0.082236 | 0.014305 |
| $1.333 < A \leq 1.5$ | 0.478382 | 0.078924 | 0.020805 |
| $1.5 < A \leq 1.667$ | 0.305372 | 0.069269 | 0.0111 |
| $1.667 < A$ | 0.298 | 0.04675 | 0.015993 |

## 6.3 Block adaptation

The magnitude of the input can be detected by buffering the input samples over a defined period of time, which we call block adaptation. Fig. 6-4 gives the simulation results of two 4th order sigma-delta modulators. One is with the fixed loop filter and the other is with the adaptive loop filter. The magnitude is detected over 16 samples at the lower rate (as in Fig. 6-2) and the adaptation logic is the same as in Table 6-1.

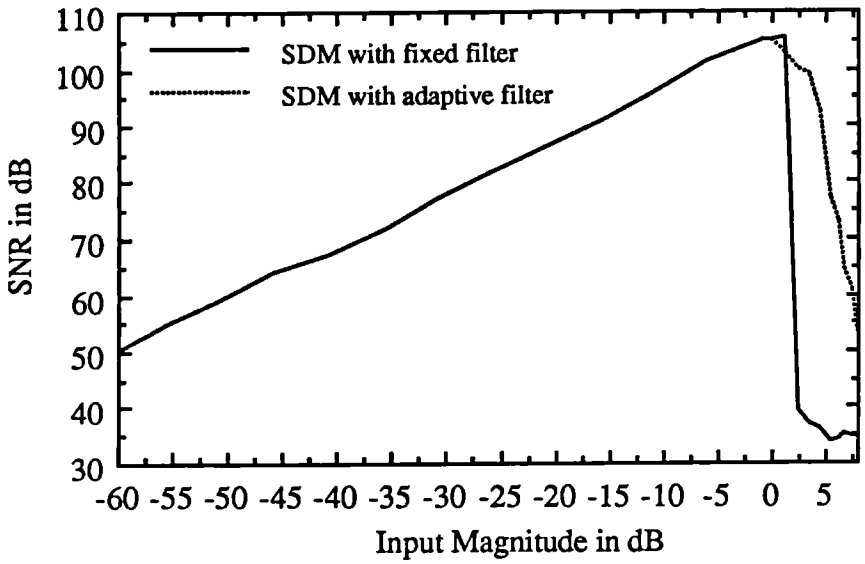



Fig. 6-4 SNR curves of SDMs with adaptive and fixed filters (block adaptation)

From Fig. 6-4, it can be seen that when input is below 0 dB, the two curves are exactly the same. However, between 0 dB and 8 dB, the one with fixed loop filter will have severe overload distortion. Although the signal-to-noise ratio will decrease for the adaptive one, no overload distortion occurs. It can be said that the dynamic range of the

system can be improved by about 5-8 dB by adapting the $\{b_i\}$ coefficients of the filter. When the magnitude reaches about 4 dB, the SNR drops to about 90 dB which a 3rd order SDM can obtain. When the magnitude increases to 6 dB, the SNR decreases to about 70 dB which is the quality of a 2nd order SDM. When the magnitude reaches about 8 dB, the SNR falls to about 54 dB and the system is equivalent to the 1st order SDM. This adaptation logic can be easily carried out by digital circuits.

The block adaptation method causes delay. In the system in Fig. 6-1, one block of input samples is stored in a buffer, and the magnitude detection is carried out over and applied to the current block. The delay will be the size of the buffer. In Fig. 6-2, the delay will also depend on the delay of the interpolator. If the two delays are quite different, the shorter one can be adjusted to the longer one by introducing an extra delay. Sometimes an error may occur because of the delay problem.

## 6.4 Instantaneous adaptation

The other way of adaptation is sample by sample. This is carried out by using the system in Fig. 6-1. In other words, the sample is taken after interpolation and the value of this sample will be directly used to decide the coefficients. There is no delay problem, but the adaptation logic has to be carried out at much higher speed. Fig. 6-5 shows the simulation results. Fig. 6-6 shows the difference in SNR between the block and instantaneous adaptation. It can be seen that there is no significant difference. The block adaptation is slightly better when input magnitude is between 6 and 8 dB.

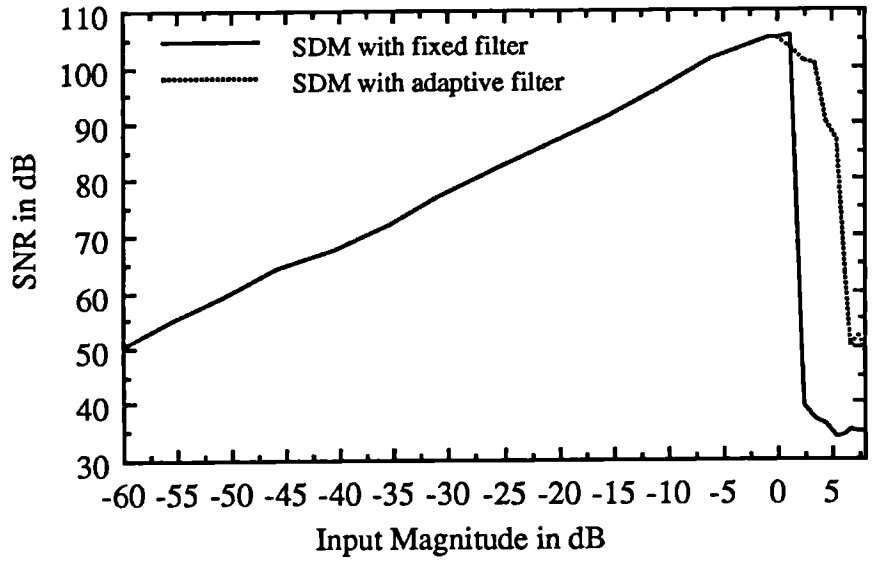



Fig. 6-5 SNR curves of SDMs with adaptive and fixed filters (instantaneous adaptation)

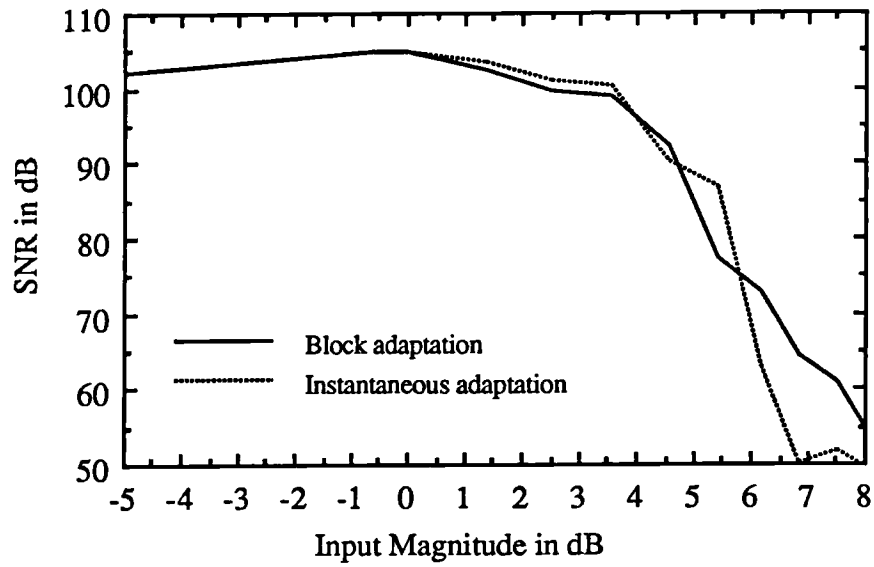



Fig. 6-6 Comparison between block and instantaneous adaptation

## 6.5 Adaptation of one coefficient

The adaptation logic may be simplified by reducing the number of adapted coefficients. It is found that if most of the coefficients are fixed and only one coefficient is optimised within the different ranges of input, the effects of different individual coefficient on the SNR are different. The changes of $b_2$ will affect the stability and the signal-to-noise ratio dramatically, i.e., if we offset $b_2$ from the point which has already been set to be optimal, it is quite likely that the system will be unstable. The computer simulations show that only varying $b_2$ cannot improve the signal-to-noise ratio radically under the large input magnitude. It is also found that the system is not very sensitive to the change of $b_4$ because it is very small. However, the optimisation of $b_3$ will improve the SNRs for different input levels. Therefore, $b_2$ and $b_4$ are fixed in adaptation process and as before, $b_1$ is always set to be one. Finally, only $b_3$ is adapted according to the input in the 4th order system. Table 6-2 gives the simulation results of optimisation of $b_3$ for different ranges of input, with $b_1$=1.0, $b_2$=0.5, and $b_4$=0.0205. Fig. 6-7 shows the SNR curves of SDM with only $b_3$ adaptation compared with its fixed SDM. In Fig. 6-8, the SNR curves of adaptations of four coefficients and only one coefficient are placed together to show that there is no big difference between them. However, adaptation of one coefficient is much simpler.

Table 6-2 Simulation results of optimal $b_3$ for different ranges of input magnitude with fixed $b_1$, $b_2$ and $b_4$

| $b_3$ | Input Magnitude A |
|---|---|
| 0.1301 | $0.0 \leq A < 1.0$ |
| 0.11 | $1.0 \leq A < 1.167$ |
| 0.09 | $1.167 \leq A < 1.333$ |
| 0.07 | $1.333 \leq A < 1.5$ |
| 0.05 | $1.5 \leq A < 1.667$ |
| 0.03 | $1.667 < A$ |

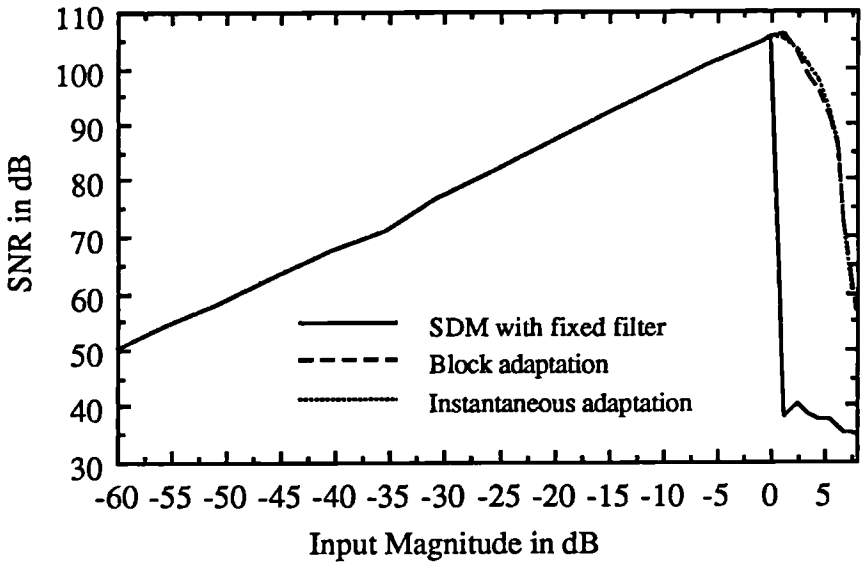



Fig. 6-7 SNR results of SDM with $b_3$ adaptation compared with the fixed SDM

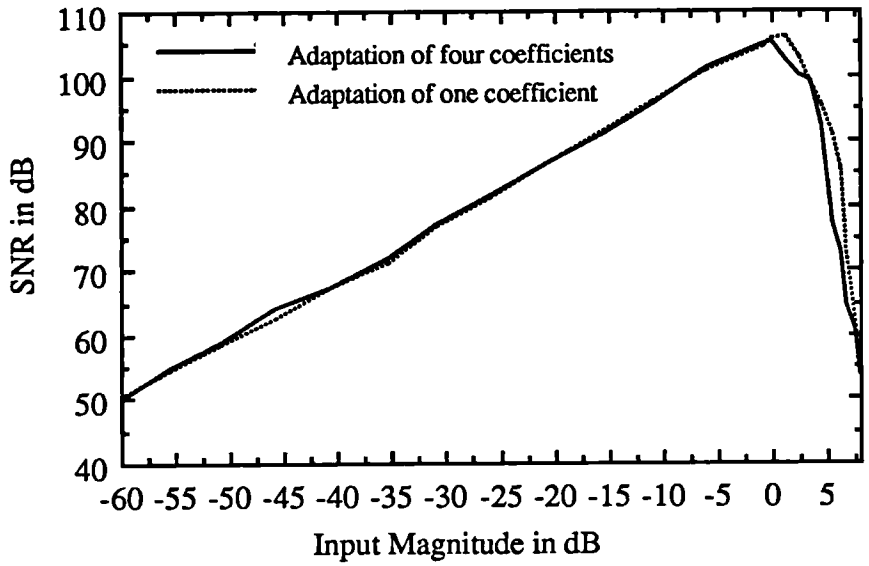
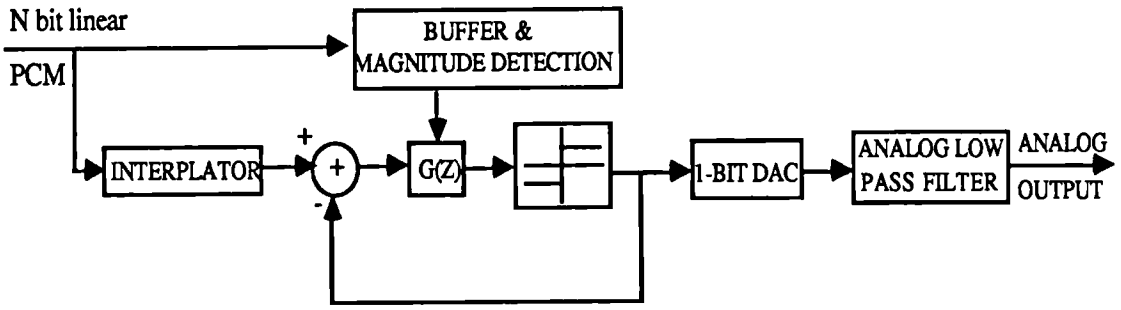


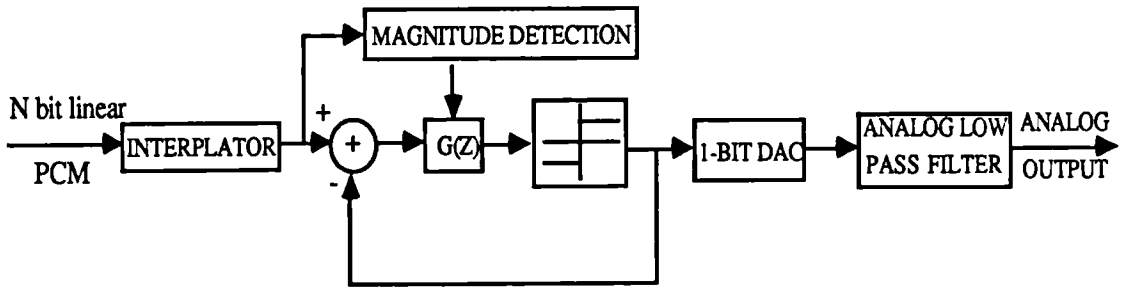

Fig.6-8 Comparison in SNR between adaptation of four coefficients and one coefficient

## 6.6 AFDM used as a D/A converter

A sigma-delta modulator with adaptive loop filter (AFDM) can be easily used as a D/A converter. The block diagrams are depicted in Fig. 6-9 for block and instantaneous logic. However, it seems more difficult to use it as an A/D converter because, firstly, it is difficult to update the coefficients of the analogue filter, and secondly, to obtain the magnitude of the input, the backward logic which is similar to the one used in adaptive quantisation in Chapter 5 has to be used. This will introduce delay and increase the complexity of the circuitry.




(a) Block logic




(b) Instantaneous logic

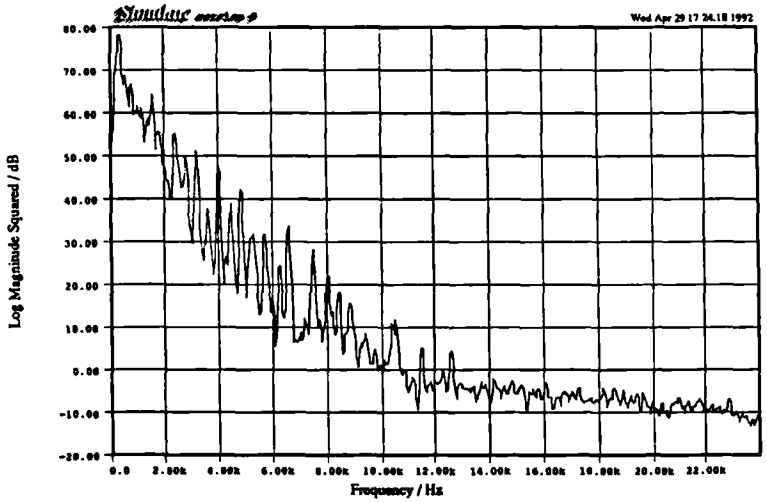
Fig. 6-9 Sigma-delta modulators with adaptive filter used as D/A converters

## 6.7 Musical signal test

As is shown in Fig. 5-10, the large magnitudes which are near to the full scale of the SDM system are relatively rare in music signals. This indicates that the system works mostly with constant coefficients, which are the cases of A<1.0 in Tables 6-1 and 6-2. Taking a 328-ms piece of music signal for the test, the maximum magnitude is set to full scale. The simulation is carried out for block adaptation of $b_3$. Among 984 blocks (16 samples/block at Nyquist rate: 48 kHz), only 24 blocks (2.4%) chose b3=0.11, 6 blocks (0.6%) for 0.09, 2 blocks (0.2%) for 0.07, and 4 blocks (0.4%) for

0.05 and 0.03. The remaining 948 blocks (96.3%) chose b3=0.1301. Fig. 6-10 shows the spectra of both the original and the reconstructed signals. Fig. 6-11 gives their time-domain waveforms over 22.7 ms.

(a)




(b)

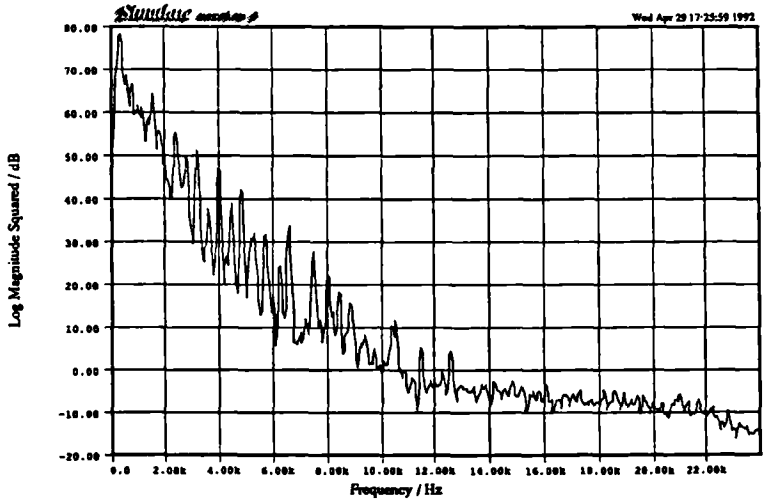
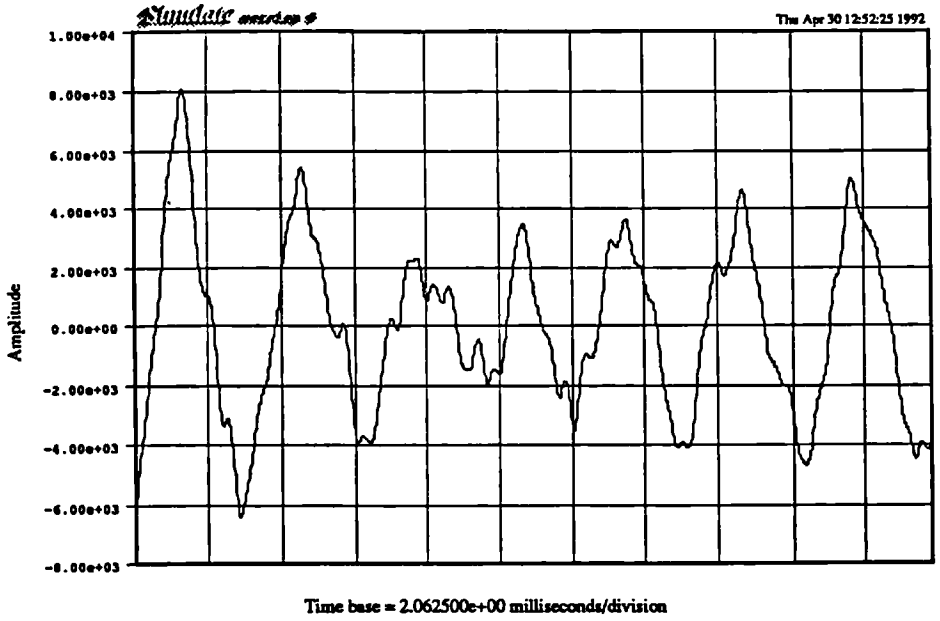
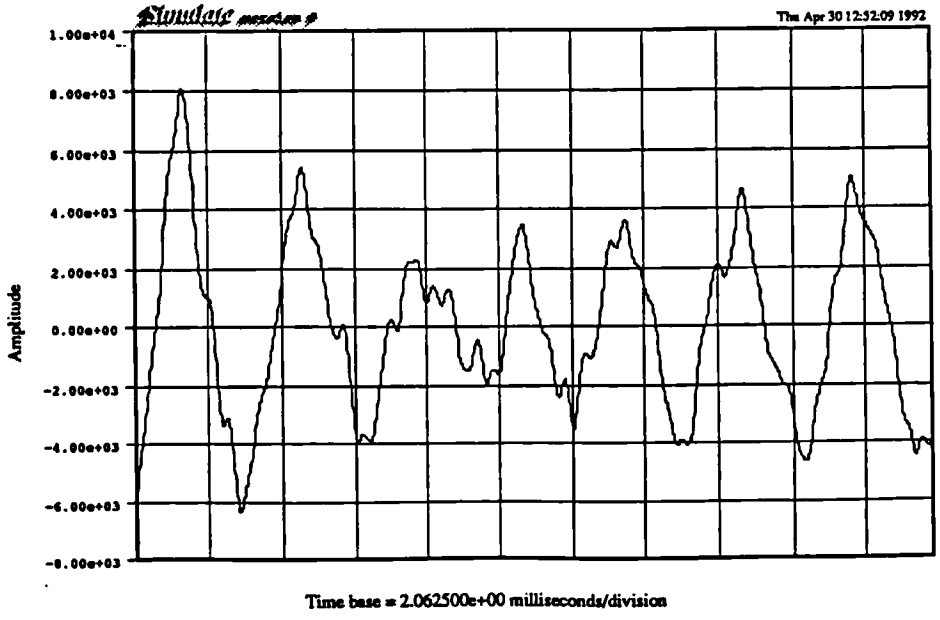



Fig. 6-10 Spectra of the original and reconstructed signals over 328 ms:
(a) original; (b) reconstructed




(a)




(b)

Fig. 6-11 Time-domain waveforms over 22.7 ms: (a) original; (b) reconstructed

## 6.8 Conclusions

In this chapter, we have investigated the possibility of adapting the loop filter in SDM systems. The stability problem of one-bit SDM makes it difficult to adapt the filter coefficients. The coefficients have to be optimised under different input ranges and then stored in a table. The adaptation is carried out by detecting the magnitude of the input and then looking up at the table. The results have shown that the dynamic range can be slightly improved by 5-8 dB by using a 4th order adaptive filter. Instead of adaptating all four coefficients, the same improvement can be achieved by only adapting one coefficient. It is easy to use an AFDM as a D/A converter but its application for A/D convertion is difficult.

# 7

# SUMMARY AND FUTURE WORK

## 7.1 Summary and discussion

This thesis has provided design methods and analyses for oversampled fixed and adaptive sigma-delta modulation (SDM) systems. The analyses and design have been carried out on fixed sigma-delta modulators. A new optimisation method for designing higher order fixed SDM systems has been proposed to overcome the conflicts which appear in traditional filter design methods for one-bit SDM. The work then has been concentrated on adaptive sigma-delta modulators. It has been shown that the dynamic range of an adaptive SDM is much wider than a fixed SDM. The author hopes that these results may be used for further research work and/or possible hardware implementation.

In Chapter 3, an optimisation method called pattern search has been used to obtain the optimal feed-forward coefficients $\{b_i\}$ and feedback coefficients $\{a_i\}$ of the loop filter in the sense of signal-to-noise ratio and stability. Once a group of optimal feed-forward coefficients $\{b_i\}$ are determined, infinite groups of coefficients can be

obtained by multiplying by any positive factors, by which the system will maintain the same performance. This is due to the nonlinearity of the one-bit SDM. The optimal quantisation level with respect to the maximum input level has been investigated by using the concept of equivalent quantiser. It has been discovered that the equivalent quantiser is time-varying and its output points are not the midpoints of the corresponding input ranges.

The stability is still a difficult topic for a fixed sigma-delta modulation system, although the system is apparently simple. In Chapter 4, an attempt has been made to discover more about it from viewpoints of nonlinearity, limit cycles and overload distortion. The phenomena of nonlinearity show that the traditional concept of noise transfer function may not be suitable for designing the loop filter. The clippers are important for preventing or alleviating the overload distortion aside from careful design of the loop filter and the quantisation level. From the property described by Chapter 3, it may be stated that the overload characteristic is more related to the relationships among the filter coefficients rather than the absolute values of them, and therefore, the clippers have to be placed after each integrator.

Chapters 5 and 6 are devoted to adaptive sigma-delta modulators. Adaptive quantiser and adaptive filter in SDM systems have been investigated separately. A feedback digital logic based on the concept of equivalent quantiser has been proposed for adaptive quantisation while a feed-forward logic based on a table-look up method is used for adaptive filter. A SDM with adaptive quantiser can improve the dynamic range dramatically. However, a SDM with adaptive filter can only slightly increase the dynamic range because of the difficulty of maintaining the system stable. Double-block calculation method is used for adaptive quantisation in order to have a fast attack but a slow release time. The adaptation is carried out over a short block and the magnitude is

detected over a long block. When using a SDM with adaptive quantisation as an A/D converter, the oversampling ratio or the order of the loop filter can be reduced while maintaining the same dynamic range compared with the fixed SDM system. By defining the minimum quantisation level appropriately, the idle channel noise can be reduced to a very low level. Adaptation logic for the loop filter can be simplified by reducing the number of adapted coefficients because some of the coefficients are not crucial in improving the SNR under a certain input level.

It may be interesting to compare the adaptive SDM with the NICAM system used by BBC since 1981[49]. The acronym NICAM stands for "Near Instantaneously Companded Audio Multiplex". The system was designed for transmitting audio on digital circuits used for multi-channel telephony. In a NICAM system, blocks of samples are examined to discover the maximum sample value in the block. The peak amplitude is then used to control a programmable gain amplifier, which adjusts the amplitude of the sampled audio. All samples in a block are coded to an accuracy determined by that largest sample value. A SDM with adaptive quantisation is quite similar to a NICAM system in some aspects. In the adaptive SDM, the peak amplitude is detected through a feedback logic and used to adjust the amplitude of the signal in the feedback path. In the NICAM system, each block contains 32 samples, which is corresponding to 1 ms at the system sampling rate of 32 kHz. In the adaptive SDM, 80 samples are included in one calculation block, which is corresponding to 1.667 ms at 48 kHz sampling frequency and 1.814 ms at 44.1 kHz sampling frequency. The difference is that the input signal is compressed and then coded in a NICAM system, but in a SDM with adaptive quantisation, the one-bit output signal is companded and then coded, and fed back to the input.

## 7.2 Future work

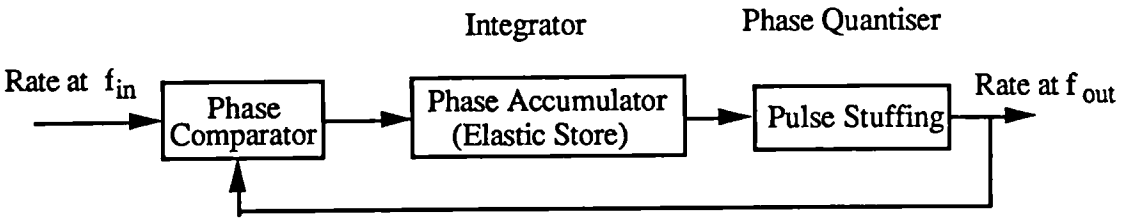
Although the research on sigma-delta modulation has been carried out for many years, there are still many open problems. One of them is the stability analysis and the use of it for designing a SDM system, especially for a single-bit case. Up to now, most of the theoretical analysis appeared in the simplest first order SDM with only one feed-forward coefficient and from the angle of limit cycles. Furthermore, the link between the analysis of limit cycles of a SDM system and the design of the system is not clear. More accurate analyses are needed for higher order SDM systems. The design work so far is mainly based on computer simulations because of the nonlinearity problem. Ideally, a theoretical frame should be provided in the design procedure and this should be at least a direction to approach.

Another open problem for the fixed SDM systems is whether the error appearing at the signal frequency in the spectrum of the final reconstructed signal is strongly related to the signal frequency or not. Steele has proved that the SNR is independent of the input frequency for a first order SDM system [25]. This is deduced by assuming that the spectral density of the noise is substantially flat over the signal band. As we have shown in the thesis, for sinusoidal inputs, there is an approximately flat output spectrum over the signal band with signal line spectrum itself. If the error at the signal frequency is strongly related to the signal frequency, then the SNR measurement in the frequency-domain is not accurate enough. This needs to be investigated extensively.

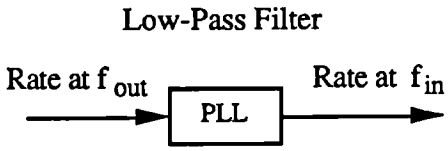
More work needs to be carried out on adaptive SDM systems. In systems with adaptive quantisation, the noise depends on the signal level, the noise being least for the lowest signal levels, and highest when the signals reach their maximum levels. The

system might produce the modulation noise. This may be very disturbing when the signal is low frequency. Although the block adaptation is better than the instantaneous one in this aspect, it is necessary to investigate the effect of the modulation noise on the SDM systems. Adaptive logic may also be extended to multi-bit or multi-stage systems. Intuitively speaking, it is easier to extend it to multi-bit SDM systems, but seems more difficult for multi-stage (MASH) systems. As we know, dithering will remove or reduce the spikes appearing in the spectrum, i.e., it will help to break certain patterns. The possibility of combining dithering technique with adaptive SDMs needs to be investigated. This may further improve the performance of the system. The dither noise may have to be designed such that it tracks the changes of the quantisation level. Although the theoretical analyses and computer simulations are essential and the first steps for adaptive SDM (ASDM) systems, the final purpose should be the real-time implementation. The future work may focus on the implementation of ASDM on DSP chips and on VLSI design of it.

It is found that the structure of a SDM system is quite similar to an asynchronous multiplexing system by using pulse-stuffing synchronisation techniques in communication systems [50]. Fig. 7-1 gives illustration diagrams of the multiplexing-demultiplexing system. In order to multiplex asynchronous digital signals, the data from different individual sources have to be synchronised at the same rate $f_{out}$. When the phase difference accumulated in the elastic store reaches a given level, the pulse stuffing event occurs. At the demultiplex, after the removal of the stuffed bits, there are gaps in the steady stream of bits. Then, a phase-locked loop (PLL) circuit can be used to smooth the data. As a result, achievements obtained in SDM systems may be applied or extended to asynchronous multiplexing systems.




(a) Multiplexing stage




(b) Demultiplexing stage

Fig. 7-1 An illustration of a multiplexing-demultiplexing system

# Appendix A

# SNR CALCULATION

The signal-to-noise ratio (SNR) is one of the important measurements for A/D, D/A conversion and coding systems. The methods of measuring SNR can be basically divided into two types: the time-domain method and the frequency-domain method.

## The time-domain method

Suppose that the reconstructed signal of SDM is $x_r(n)$ and divided into two parts

$$x_r(n) = x(n) + e(n)$$

where x(n) is the original signal and e(n) is the error produced by SDM. A standard objective measure of quality is the ratio of signal variance to error variance, referred to for historical reasons as the signal-to-noise ratio (SNR):

$$SNR = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E[x^2(n)]}{E[e^2(n)]} \qquad \text{(A-1)}$$

where $\sigma_x^2$ and $\sigma_e^2$ are the variances of signal and error respectively. Note we are assuming that both x(n) and e(n) have zero mean. If this is not the case, the mean value of them should be subtracted out prior to SNR calculations. (A-1) is one of the most utilised descriptor of performance of coding system.

One of the estimates for the variance is the sample variance defined as [51]

$$\hat{\sigma}_x^2 = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \qquad \hat{\sigma}_e^2 = \frac{1}{N}\sum_{n=0}^{N-1} e^2(n) \tag{A-2}$$

If the processes are Gaussian processes, the estimates will be maximum likelihood. If they are ergodic processes, then, when N approaches infinity, the time averages equal ensemble averages, i.e.,

$$\hat{\sigma}_x^2 = \sigma_x^2 \quad \text{and} \quad \hat{\sigma}_e^2 = \sigma_e^2$$

By using the estimates of (A-2), (A-1) becomes

$$\text{SNR} = \frac{\sum_n x^2(n)}{\sum_n e^2(n)} \tag{A-3}$$

(A-3) is a time-domain estimation of SNR, where e(n) can be obtained by

$$e(n) = x_r(n) - x(n)$$

Based on this time-domain measurement, several other methods have been developed for different purposes and situations. For example, segmental signal-to-noise ratio (SNRSEG) emphasises the weak-signal performance, which is often used in speech processing.

The basic time-domain method (A-3) is easy to carry out in computer simulations. The disadvantage of this method is that we need to know the exact delay and gain of the measured system. As we know, the gain and delay will not introduce noise on the signal. However, if they cannot be determined correctly, the result of SNR calculation may be very poor, which does not reflect the real SNR.

Another method is called sinusoidal minimum error method [52], which uses sinusoids as test signals. Sinusoidal signals are easy to define and generate so that they are widely used for system measurements [53]. For a sinusoidal input

$$x(nT) = A\cos(2\pi f_x nT)$$

the output of an A/D or D/A conversion system is a sinusoidal signal at the input frequency, $f_x$, together with the error introduced by the quantisation. The output $x_r(nT)$ can be divided into two parts: signal-correlated part $x_x(nT)$, and signal-uncorrelated part $e_u(nT)$:

$$x_r(nT) = x_x(nT) + e_u(nT)$$

where

$$x_x(nT) = a_0 + a_1\cos(2\pi f_x nT+\phi_1) + \sum_{k=2}^{\infty} a_k\cos(2\pi k f_x nT + \phi_k)$$

where the first term $a_0$ is offset, the second term is signal component and the remaining parts are harmonics.

The signal power at the output is

$$P_{out} = \frac{a_1^2}{2}$$

and the power in the harmonics is

$$P_h = \frac{1}{2}\sum_{k=2}^{\infty} a_k^2$$

Suppose $P_e$ is the power of $e_u(nT)$, then, the total SNR is

$$\text{SNR} = \frac{P_{out}}{P_e + P_h} \qquad \text{(A-4)}$$

The linear properties of an A/D converter can be characterised by its gain, $G=a_1/A$ and phase $\phi_1$, both possibly functions of the signal frequency $f_x$. If $a_1$ is the function of the signal frequency, the SNR in (A-4) can reflect the change of $a_1$ by varying the

frequency. However, if $\phi_1$ is not a linear function of the frequency, i.e., the group delay is not constant along the frequency axis, then (A-4) cannot reflect the phase distortion.

A template consisting of a sinusoid at the input frequency, a dc offset term, and harmonics is employed to match $x_x(nT)$. The amplitudes, $a_k$ and phases, $\phi_k$, of the output signal and its harmonics can be determined by fitting the template to the system output $x_r(nT)$ so as to minimise the mean square (power) of the error $e_u(nT)$.

## The frequency-domain method

By using Parseval's Relation, the following hold [51]

$$\sum_{n=0}^{N-1} x^2(n) = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2 \qquad \sum_{n=0}^{N-1} e^2(n) = \frac{1}{N} \sum_{k=0}^{N-1} |E(k)|^2$$

where X(k) and E(k) are N-point DFT of x(n) and e(n). Therefore, (A-3) becomes

$$SNR = \frac{\sum_{k=0}^{N-1} |X(k)|^2}{\sum_{k=0}^{N-1} |E(k)|^2}$$

SNR is usually expressed in decibels (dB):

$$SNR\ (dB) = 10 \log_{10} \left( \frac{\sum_{k=0}^{N-1} |X(k)|^2}{\sum_{k=0}^{N-1} |E(k)|^2} \right)$$

In the case of single-tone sinusoidal input, the signal is within a very narrow band. That is, $X(k) \approx 0$ for most of k. Suppose that $X(k) \neq 0$ when $k_1 \leq k \leq k_2$, where $k_2$-$k_1$

is very small compared to N. If E(k) is much smaller compared with the signal in the narrow signal band, then the noise within this narrow band can be ignored, i.e., E(k) ≈ 0 when $k_1 \le k \le k_2$. Thus, the SNR can be estimated by only calculating the spectrum of reconstructed signal $x_r(n)$. Also considering the symmetric property of the DFT, SNR is therefore approximately

$$SNR \approx \frac{\sum_{k=k_1}^{k_2} |X_r(k)|^2}{\sum_{k=0}^{k_1-1} |X_r(k)|^2 + \sum_{k=k_2+1}^{\frac{N}{2}-1} |X_r(k)|^2}$$

Fig. A-1 illustrates the calculation method. The components outside ($k_1$, $k_2$) include the signal-uncorrelated noise and harmonic distortion. The test frequency must be chosen so that harmonics aliased into the baseband do not add to the fundamental. The data from the system should be modified by a window to reduce the effects of truncating before calculating FFT.

The advantage of the frequency-domain method is that we do not need to consider gain and delay of the system, which is the same as that of the sinusoidal minimum error method. However, in computer simulations, the DFT of the signal has to be calculated. In practice, it requires the use of a narrow-band rejection filter in the receiver equipment to block the sinusoidal test signal from the distortion measuring circuits so that the distortion power can be measured. The details can be found in the CCITT standards [39][53]. This method requires much more calculations than the sinusoidal minimum error method. Nevertheless, based on the existing FFT spectral analysis, it calls for very few extra calculations. The frequency-domain method has been tested using computer simulations by measuring PCM systems with different bit numbers. The

differences are less than half dB between the calculated results and the theoretical values. The big disadvantage of the method is that it cannot be used for test signals other than sinusoids. Furthermore, if the system has severe phase distortion, i.e., the group delay of the system is not constant, this method will introduce error.

In systems like sigma-delta modulators, the interpolators and decimators are included, which often introduce noninteger delay in the discrete-time axis. Therefore, it is more convenient to use frequency-domain method, or sinusoidal minimum error method.

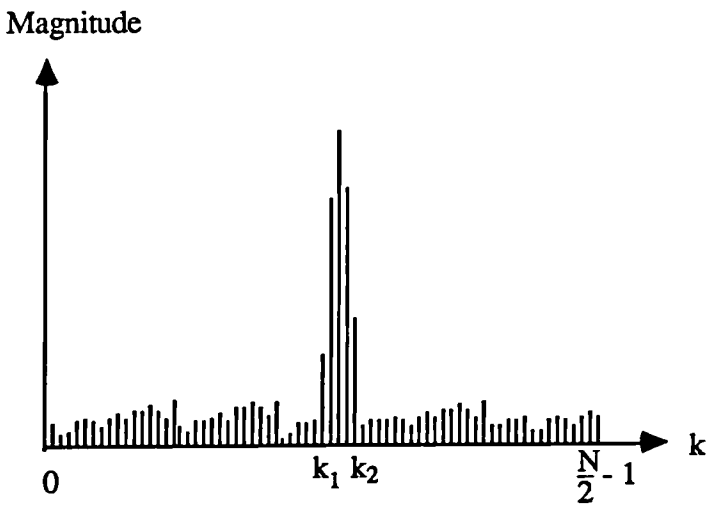



Fig. A-1 Illustration of SNR calculation in the frequency-domain

# Appendix B

# OPTIMISATION METHOD FOR DESIGNING SDM

The requirement for methods of optimisation arises from the mathematical complexity necessary to describe the theory of systems, processes, equipment, and devices which occur in practice. Even quite simple systems must sometimes be represented by theory which may contain approximations, by parameters which change with time, or by parameters that vary in a random manner. For many reasons the theory is imperfect, yet it must be used to predict the optimum operating conditions of a system such that some performance criterion is satisfied. At best such theory can predict only that the system is near to the desired optimum. Optimisation methods are then used to explore the local region of operation and predict the way that the system parameters should be adjusted to bring the system to optimum.

There are usually two different optimisation problems: minimisation and maximisation. They can be merged into one by the following:

$E_1$ = maximum f(x) $\Leftrightarrow$ $E_2$ = minimum {-f(x)}

or $E_1$ = maximum f(x) $\Leftrightarrow$ $E_2$ = minimum {1/f(x)}

The function f is referred to as the objective function whose value is the quantity which is to be minimised or maximised.

Methods for multi-variable optimisation fall into two classes: search methods which use function evaluation only, and gradient methods which in addition require gradient information, although those classes are not completely separate. In the optimisation process of a SDM system, the gradient information is not available so that the search methods have to be chosen.

Search methods for optimisation attempt to reduce the value $E_2$ or increase the value $E_1$ of the objective function f by the use of tests near to an estimate of the solution. The tests determine a direction of search in which the minimum or maximum is expected to lie. The optimum is then approached by taking a fixed step towards it. Since the direction determined is not necessarily correct, the process is iterative. After each optimisation step, further searches are carried out until a criterion, which defines when the desired optimum has been found, is satisfied.

One of the search methods is pattern search. The basic pattern search takes incremental steps after suitable directions have been found by local exploration. If the search progresses well, the step size is increased. If it is not progressing, either because the optimum is near or because of difficulties (e.g. a narrow valley), the step size is reduced. Therefore, this method can automatically adjust the step size according to the situation. When the step size is reduced below a set figure the search is ended. The detail of this method is described in [54].

Considering a SDM system, the objective function is SNR which has the coefficients $\{a_i\}$ and $\{b_i\}$ of the loop filter in Fig. 3-3 as its variables. The purpose of optimisation is to maximise SNR. It has been found that when optimising $\{b_i\}$ coefficients, there are many local optimum points. Some of them result in nearly the

same SNR values which it is reasonable to consider as the final optimum results for a certain SDM. But some of them cause very poor SNR. The reason is that the searching process is trapped in a small neighbourhood in which a local optimum, which is not good enough for the system, is located. In order to avoid this case, an additional SNR test is added to the basic pattern search. If the local optimum is lower than a set figure, then change the initial guess and start the search from the beginning again. The flowchart of the programme is shown in Fig. B-1.

In Fig. B-1, the SNR (signal-to-noise ratio) is calculated by calling a subroutine. This subroutine includes a simulator of sigma-delta modulator, decimation filter, FFT programme, and the function for calculating the SNR of the output. The flowchart of this subroutine is shown in Fig. B-2.

Consider n coefficients which need to be optimised as a vector of n dimension. The step size in each direction can be set differently. A lower bound is preset for each step size. If all the step sizes in each direction are smaller than their lower bounds (call them step_bound), the iteration progress ends. The following are some examples.

<u>The first order SDM</u>

Condition: $b_1$=1.0

Initialisations: $a_1$=0.0, step_size=0.005, step_bound=3.91e-05

Results: iteration times=13, $a_1$= -5.9375e-03

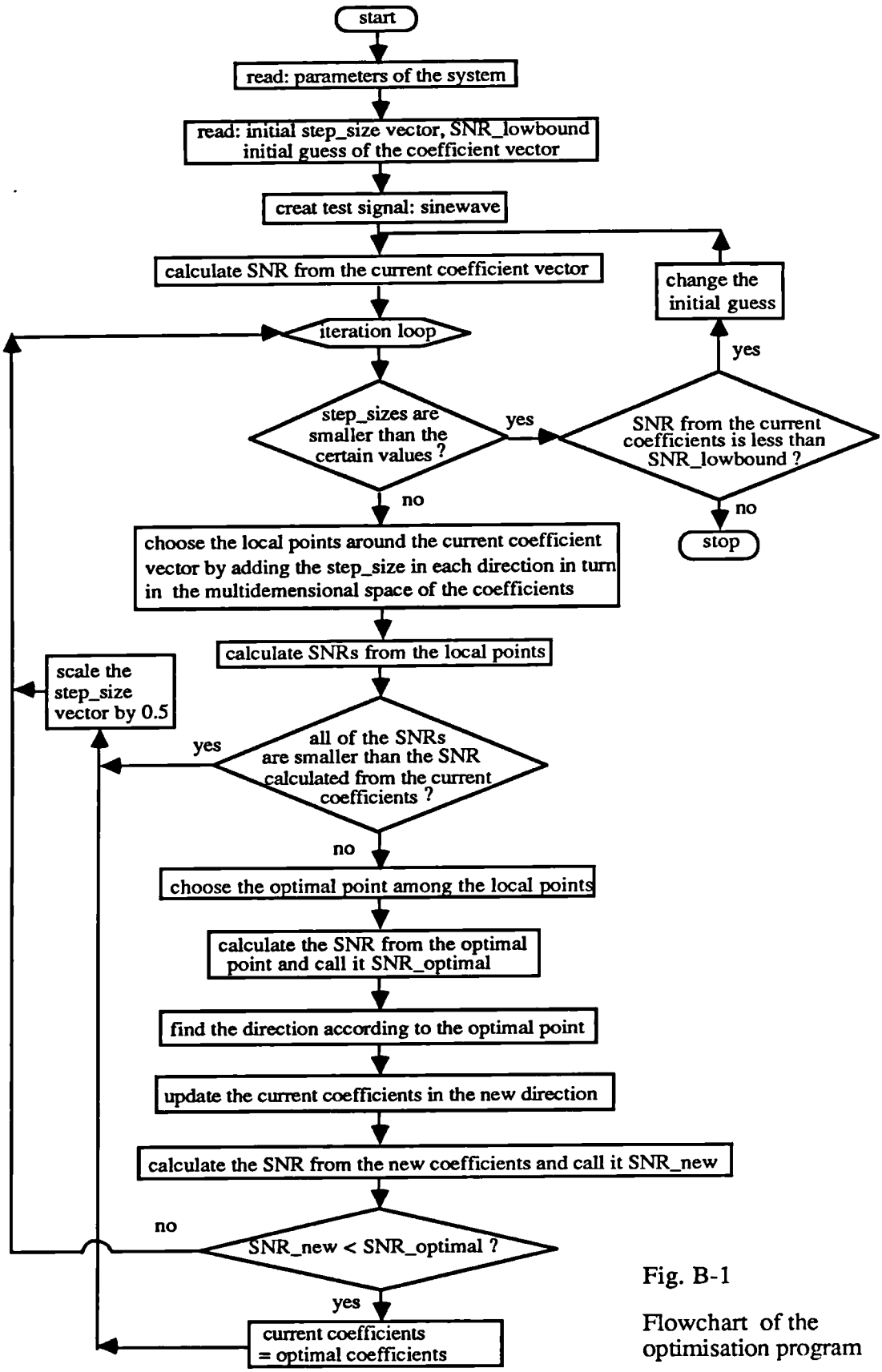



Fig. B-1

Flowchart of the optimisation program

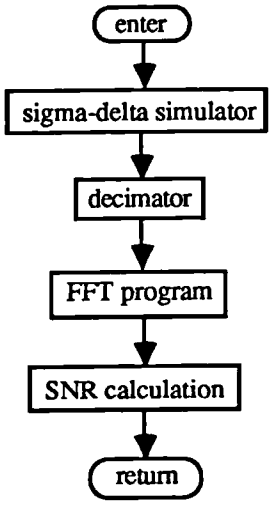



Fig. B-2 Flowchart of the subroutine for calculating the SNR

The second order SDM

(1) $b_i$ coefficients

Conditions: $b_1$=1.0, $a_1$=$a_2$=0.0

Initialisations: $b_2$=0.56, step_size=0.04, step_bound=6.25e-04

Results: iteration times=14, $b_2$=0.395625

(2) $a_i$ coefficients

Conditions: $b_1$=1.0, $b_2$= 0.395625

Initialisations: $a_1$=0.01, $a_2$=0.0, step_sizes = 0.005, 0.001,

step_bounds=1.5625e-04, 3.125e-05

Results: iteration times=13, $a_1$=2.8125e-03, $a_2$=7.5e-04

The third order SDM

(1) $b_i$ coefficients

Conditions: $b_1$=1.0, $a_1$=$a_2$=$a_3$=0.0

Initialisations: $b_2$=0.435, $b_3$=0.142, step_sizes=0.2, 0.1,

step_bounds=6.103e-06, 3.051e-06

Results: iteration times=25, $b_2$=0.447112, $b_3$=0.140976

(2) $a_i$ coefficients

Conditions: $b_1$=1.0, $b_2$=0.447112, $b_3$=0.140976

Initialisations: $a_1$=$a_2$=$a_3$= 0.0, step_sizes = 0.01, 0.005, 0.001,

step_bounds=3.05e-07, 1.52e-07, 3.05e-08

Results: iteration times=31, $a_1$=0.002, $a_2$=1.2187e-03, $a_3$=1.3458e-06

The fourth order SDM

(1) $b_i$ coefficients

Conditions: $b_1$=1.0, $a_1$=$a_2$=$a_3$=$a_4$=0.0

Initialisations: $b_2$=0.49668, $b_3$=0.08387, $b_4$=0.049355, step_sizes=0.1, 0.05, 0.001,

step_bounds=3.051e-06, 1.525e-06, 7.62e-07

Results: iteration times=25, $b_2$=0.5459, $b_3$=0.133846, $b_4$=0.022432

(2) $a_i$ coefficients

Conditions: $b_1$=1.0, $b_2$=0.5459, $b_3$=0.133846, $b_4$=0.022432

Initialisations: $a_1$=1.0e-04,$a_2$=1.0e-05, $a_3$=1.0e-06, $a_4$=0.0,

step_sizes = 0.004, 0.001, 0.0007, 0.0005

step_bounds=1.953e-06, 4.88e-07, 3.41e-07, 2.44e-07

Results: iteration times=29, $a_1$=8.564844e-03, $a_2$=1.51e-03, $a_3$=1.0e-06, $a_4$=0.0

# Appendix C

# TRANSFORMATION FOR CALCULATING LIMIT CYCLES

For convenience, the first order sigma-delta modulator in Fig. 2-10 is depicted in Fig.C-1 again with the quantisation level being one ( normalised case ). Throughout this appendix, the input is assumed to be a constant or dc value and $x \in (-1, 1]$. For this case the integrator state variable $u_i$ is given by the recursion relation:

$$u_i = x - q(u_{i-1}) + u_{i-1} \qquad \text{(C-1)}$$

Let $\gamma \equiv (1-x)/2$ and $w_i \equiv u_i/2 + \gamma$. Then (C-1) becomes

$$u_i = 1 - 2\gamma + u_{i-1} - q(u_{i-1})$$

$$= 1 - q(u_{i-1}) - 2\gamma + 2(w_{i-1} - \gamma).$$

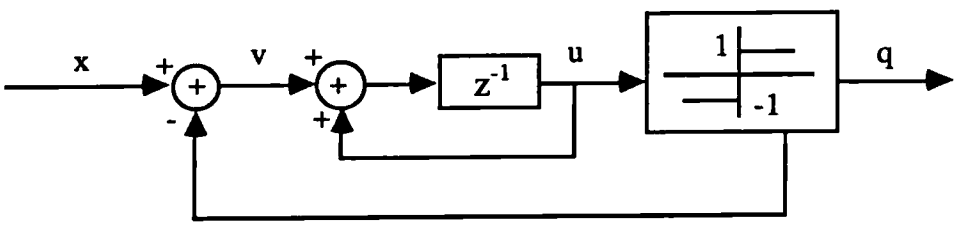



Fig. C-1 First order one-bit sigma-delta modulator

It is not difficult to see that $\gamma \in [0, 1)$. If $u_0 \in [x-1, x+1)$ then $u_i \in [x-1, x+1)$ for every i [10] and therefore $w_i \in [(x-1)/2+\gamma, (x+1)/2+\gamma) = [0, 1)$. Then $w_i$ satisfies the nonlinear recursion

$$w_i = (w_{i-1} - \gamma) + (1 - q(w_{i-1} - \gamma))/2$$

$$= \langle w_{i-1} - \gamma \rangle$$

$$= Tw_{i-1}$$

where $\langle x \rangle = x \bmod 1$

$$= \begin{cases} 1+x, & x < 0 \\ x, & x > 0 \end{cases}$$

and $T : [0, 1) \rightarrow [0, 1)$ is a transformation defined by

$$Tw = \langle w - \gamma \rangle .$$

Fig.C-2 is the graphical representation of T with $\gamma$=0.7.

Let $w_0 = u_0/2 + \gamma$. Then $w_i = T^i w_0$. Using this relation we can obtain the sequence $u_i$ by investigating the trajectory of $w_0$ under repeated application of the transformation T. Let D be defined by $D = 1 - \gamma$. Then

$$T^0D = D = 1 - \gamma$$

$$T^1D = \langle T^0D - \gamma \rangle = \langle 1 - 2\gamma \rangle = \begin{cases} 2 - 2\gamma, & T^0D < \gamma \\ 1 - 2\gamma, & T^0D > \gamma \end{cases}$$

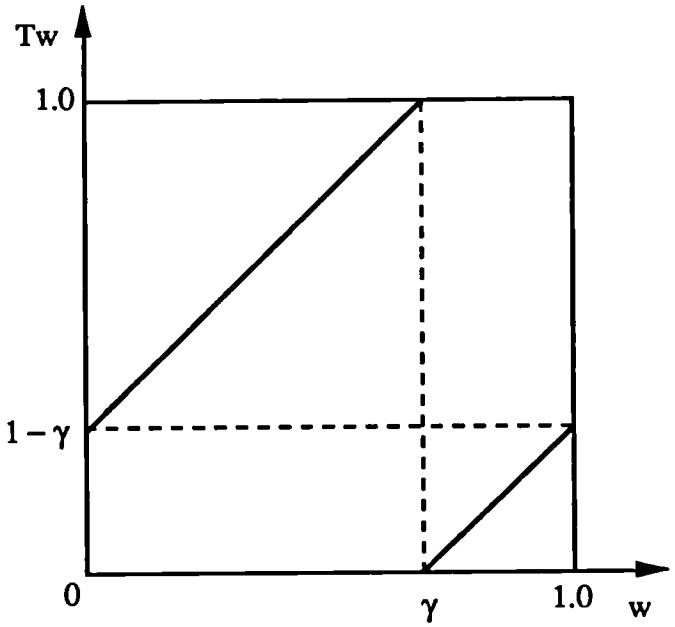



Fig. C-2 Transformation: T

$$T^2D = < T^1D - \gamma > = \begin{cases} < 2 - 3\gamma > = \begin{cases} 3 - 3\gamma, & T^1D < \gamma, \; T^0D < \gamma \\ 2 - 3\gamma, & T^1D > \gamma, \; T^0D < \gamma \end{cases} \\ < 1 - 3\gamma > = \begin{cases} 2 - 3\gamma, & T^1D < \gamma, \; T^0D > \gamma \\ 1 - 3\gamma, & T^1D > \gamma, \; T^0D > \gamma \end{cases} \end{cases}$$

$$\vdots$$

$$T^iD = 1 - (i+1)\gamma + J_i \qquad \text{(C–2)}$$

where $J_i$ depends on the value of $T^jD$ ( j = 0, 1, ... i-1 ). If $T^jD < \gamma$ for any j, one will be added to $J_i$. Furthermore, if $T^jD=\gamma$ for any j, then $T\gamma = < \gamma - \gamma > = < 0 > = 0$, or 1 and the next transformation will lead to 1-γ, which come back to the initial state D=1-γ. This indicates the periodic property of the system. Let K be the smallest integer such

that $T^K D=\gamma$, then $T^{K+2}D$ will be D so that the period is K+2. Using (C-2) when i = K, it can be obtained that

$$T^K D = 1 - (K+1)\gamma + J_K = \gamma$$

so that

$$\gamma = (1+J_K)/(K+2) \qquad \text{(C-3)}$$

For any dc input x which can be expressed as a rational number b/a, $\gamma$ can be given as

$$\gamma = (1-x)/2 = (a-b)/2a$$

so that $\gamma$ can also be expressed as a rational number because a and b are integers. Suppose that $\gamma=m/n$, where m and n are relatively prime, so that

$$\gamma n = m \rightarrow \gamma(n-2+1) + \gamma = m-1+1 \rightarrow 1 - \gamma(n-2+1) + m-1 = \gamma \qquad \text{(C-4)}$$

Considering (C-2), let i = n-2, then

$$T^{n-2}D = 1 - \gamma(n-1) + J_{n-2} \qquad \text{(C-5)}$$

From (C-4) we know that

$$1 - \gamma(n-1) = \gamma + 1-m \qquad \text{(C-6)}$$

Substituting $1 - \gamma(n-1)$ in (C-5) with (C-6), (C-5) becomes

$$T^{n-2}D = \gamma + 1-m + J_{n-2} = \gamma + B$$

where B is an integer and equal to $(J_{n-2} + 1-m)$. Considering that $\gamma \in [0,1)$, if $B>0$, then $B+\gamma>1$, and if $B<0$, then $B+\gamma<0$. As we know, $T^{n-2}D \in [0,1)$, so that the above results are conflict with the transformation property. Therefore, B must be zero, which leads to

$$J_{n-2} = m-1, \quad \text{and} \quad T^{n-2}D=\gamma. \tag{C-7}$$

From (C-3) and (C-4) we know that the following must hold:

$$\frac{1+J_K}{K+2} = \frac{m}{n} \tag{C-8}$$

As is assumed, K is the smallest integer such that $T^KD=\gamma$. Thus, from (C-7), n-2 must be greater than or equal to K, that is

$$K+2 \leq n$$

If K+2 <n, then, $1+J_K<m$. These are impossible if m and n are relatively prime so that there must be: K+2=n. This proves that in (C-8) $1+J_K$ and K+2 are also relatively prime. As the final result, the input x can be expressed as

$$x = 1 - 2m/n$$

where m and n are relatively prime and n is the period of the limit cycle in the system in Fig. C-1.

# Appendix D

# PUBLISHED WORKS

This publication was based on the author's work in Chapter 5 in this thesis.

*APPENDIX D*

# Adaptive quantisation for one-bit sigma-delta modulation

J. Yu, BSc, MSc
M.B. Sandler, BSc, PhD, CEng, MAES, MIEE
R.E. Hawken, BSc(Eng)

*Indexing terms: Adaptive quantisation, Sigma-delta modulation*

**Abstract:** A fixed step size is usually used for a quantiser in a sigma-delta modulator or noise shaper, but it cannot always match input signals adequately if they are nonstationary, as in the case of music. An attempt at introducing adaptive quantisers, based on a digital maximum-magnitude technique, into 1-bit sigma-delta modulators has been made, although the basic idea appeared about two decades ago. The initial results show it to be a promising technique. The dynamic range of the sigma-delta modulator can be effectively increased by using an adaptive quantiser, and the signal/noise ratio is nearly independent of input level for sinewave inputs. This advantage may increase future applications of sigma-delta modulators.

## 1 Introduction

Oversampled sigma-delta modulation is becoming increasingly popular, especially for A D and D/A conversion. Compared with traditional PCM, it employs considerably fewer bits for the quantiser by using oversampling and noise shaping techniques. It eliminates the need for precise analogue pre- and post-filters owing to a high oversampling ratio. It is more robust against circuit imperfections than the standard successive approximation because the single bit quantiser is less sensitive to level shifts than the multiple thresholds involved in the traditional implementations. Also, compared with oversampled delta modulation, there is no error accumulation in the decoder because there is no feedback loop. Such oversampled convertors usually require high timing accuracy, but precision in the circuit elements can be relaxed. Additional benefits can be gained with some digital processing on-chip after digitisation such as equalisation, echo cancellation, etc. Thus it can provide integrated analogue and digital functions and is well suited to VLSI implementations [1–7].

As is well known, the characteristics of the quantiser are very important. Usually, quantisers with fixed step size are used for sigma-delta modulators and these cannot always match the variance of the input. For small signal magnitude, very coarse quantisation could occur and, for a large signal, overload may occur. Adaptive quantisers have been used in digital coding systems such as APCM, ADPCM and ADM [8]. They have also been examined in some simple sigma-delta modulators (SDMs) [9, 10]. They have shown considerable advantages in increasing the dynamic range and attaining better quality at the same bit rate or reducing the bit rate of the system while maintaining the same quality.

Paper 8425G (E8, E10), first received 14th March and is revised from 16th July 1991

The authors are with the Department of Electronic and Electrical Engineering, Kings College London, University of London, Strand, London WC2R 2LS, United Kingdom

## 2 Sigma-delta modulator and noise shaper

The basic block diagrams of the digital sigma-delta modulator and demodulator are shown in Fig. 1, where

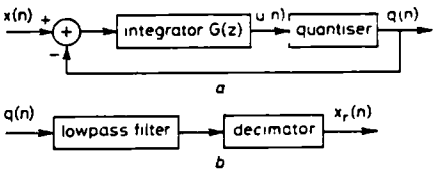



**Fig. 1** *Basic block diagram of sigma-delta modulation*
*a* Modulator
*b* Demodulator

$x(n)$ is an oversampled signal, sampled at $f_s$, and $x_r(n)$ is a reconstructed signal sampled at or slightly in excess of the Nyquist sampling frequency $f_b$. It is assumed that the wanted signal only occupies the frequency band from 0 to $f_b/2$. If the oversampling ratio is $N$, then $f_s = Nf_b$. From Fig. 1*a*, the system can be described as

$$[X(z) - Q(z)]G(z) = U(z) \tag{1}$$

Let

$$u(n) = q(n) - e(n)$$

where $e(n)$ is the quantisation noise, so that

$$U(z) = Q(z) - E(z) \tag{2}$$

Combining eqns. 1 and 2

$$Q(z) = \frac{G(z)}{1 + G(z)} X(z) + \frac{1}{1 + G(z)} E(z) \tag{3}$$

This can be generalised to

$$Q(z) = F_X(z)X(z) + F_E(z)E(z)$$

where $F_X(z)$ and $F_E(z)$ are the signal and noise transfer function, respectively

$$F_X(z) = \frac{G(z)}{1 + G(z)} \quad F_E(z) = \frac{1}{1 + G(z)}$$

For the $n$th order sigma-delta modulator, the most commonly used function $G(z)$ is (see [11])

$$G(z) = \frac{1}{(1 - z^{-1})^n} - 1$$

Therefore, eqn. 3 becomes

$$Q(z) = [1 - (1 - z^{-1})^n]X(z) + (1 - z^{-1})^n E(z)$$

that is

$$F_X(z) = 1 - (1 - z^{-1})^n$$

$$F_E(z) = (1 - z^{-1})^n \quad (4)$$

From eqn. 4, it can be seen that when the frequency $f$ is much less than $f_s$, $|F_X(e^{j2\pi f})|$ is approximately equal to one. This is the case for a baseband signal in a highly oversampled system. It can also be seen that the quantisation noise has been shaped. $F_E(z)$ behaves like a high-pass filter or differentiator. The noise is suppressed at low frequencies and amplified at higher frequencies so that the resolution in the signal band can be increased. In the demodulator, the quantised signal $q(n)$ is fed to a lowpass filter to remove out-of-band noise, and the output of the filter is decimated, i.e. it has its sampling rate reduced, to obtain the reconstructed in-band signal.

A similar kind of system is called a noise shaper; its basic block diagram is shown in Fig. 2. It can be derived

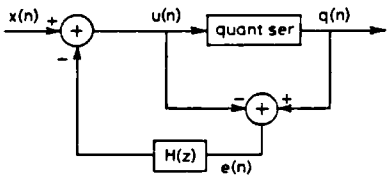



**Fig. 2** *Basic diagram of noise shaper*

from Fig. 2 that

$$Q(z) = X(z) + (1 - H(z))E(z)$$

Usually, for the $n$th-order noise shaper

$$H(z) = 1 - (1 - z^{-1})^n$$

so that

$$Q(z) = X(z) + (1 - z^{-1})^n E(z)$$

Therefore,

$$F_X(z) = 1 \quad F_E(z) = (1 - z^{-1})^n \quad (5)$$

On comparing eqns. 4 and 5, it can be seen that the noise is reshaped in the same manner. However, these two systems are slightly different with respect to the input signal , in that the in-band gain is identically one for the noise shaper but only approximately so for a SDM.

To date, a fixed step size is usually used for the quantiser in a sigma-delta modulator or noise shaper; i.e. no matter how the variance of the input changes, the step size is a constant. For a stationary input, this type of quantiser can work very well as long as the step size matches the variance of the input. But for a nonstationary input, for instance, a music signal, which usually has a large dynamic range and changing variance, the step size does not always match it effectively.

## 3 Adaptive quantisation

The magnitude of a music signal can vary over a wide range depending on the instruments, singers etc. In the digitising process, on the one hand we wish to choose the quantisation step size large enough to accommodate the maximum peak-to-peak range of the signal; but on the other hand we would like to make the quantisation step small so as to minimise the quantisation noise. It is impossible to satisfy both objectives when using a fixed quantiser. The basic idea of adaptive quantisation is to let the step size vary so as to match the variance of the input signal.

In order to adapt the step size, it is necessary to obtain an estimate of the time varying amplitude properties of the input signal. Usually, there are two types of method: forward and backward adaptation [12]. Forward adaptation is based on the estimation of unquantised samples, i.e. usually at the input of the quantiser. Backward adaptation is based on the estimation of the output of the quantiser. Fig. 3 shows their block diagrams. Forward estimates of step size are unaffected by quantisation noise, they are therefore more reliable. However, the system needs to transmit this additional information to the receiver. Although the backward estimates are not as accurate as the forward estimates, additional bits are not needed for the estimation.

A common approach to variance calculation is to assume that the variance is proportional to the short-time energy, which is defined as the output of a lowpass filter with input, $x^2(i)$ [8]. That is

$$\sigma_x^2(i) = \sum_m x^2(m)h(i - m)$$

where $h(n)$ is the impulse response of the lowpass filter.

If a 1-bit quantisation function is defined by

$$q(n) = \begin{cases} d & \text{if } u(n) \geq 0 \\ -d & \text{otherwise} \end{cases}$$

where $u(n)$ is the input of the quantiser, then the adaptive logic can be forward:

$$d_i = c\sigma_u^2(i)$$

or backward:

$$d_i = c\sigma_q^2(i)$$

where $c$ is a scaling constant, and $\sigma_u^2$ and $\sigma_q^2$ are the variance of the signal $u(n)$ and $q(n)$, respectively. It is obvious that the backward logic is always constant so it cannot be used for the one-bit case.

A large amount of calculation is needed to obtain the short-time energy. An alternative approach is to use local

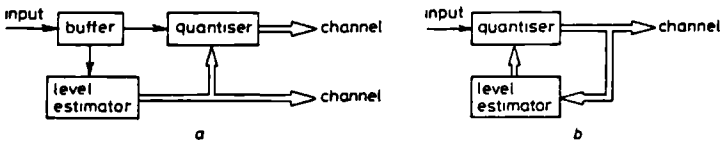



**Fig. 3** *Forward and backward adaptation*
*a* Forward adaptation *b* Backward adaptation

values of peak output magnitude to vary the overload level. Thus for a forward adaptation of a 1-bit quantiser, the logic can be in the form

$$d_i = c\,|u|_{max} \quad |u|_{max} = \max\{|u(i-k)|\} \quad k = 1, 2, \ldots, K$$

where $d_i$ is the 1-bit quantisation level of the $i$th data block and each block consists of $K$ samples. Backward adaptation based on the maximum magnitude of the output of a 1-bit quantiser is also meaningless because the maximum is always the same as the quantisation level. The maximum-magnitude logic is simpler than the variance estimation and particularly appropriate to the control of overload distortion.

Considering the case of sigma-delta modulation, the additional bits to be stored or transmitted are unwanted, so backward adaptation has to be chosen. However, backward adaptation based on the output of the quantiser is impossible owing to the characteristic of single bit quantiser. In this paper, a backward logic which is not directly based on the output of the quantiser is introduced. An estimate of the maximum magnitude of the input is used for the adaptation and will be described in detail in the following section.

## 4 Logic design of adaptation

As mentioned above, adaptive quantisation based on the maximum-magnitude logic is easier than that based on the variance estimation. If the step size of the quantiser can be changed as the maximum magnitude of the input $u(n)$ to the quantiser changes, the dynamic range of the sigma-delta modulator can be increased. However, it is difficult to estimate the magnitude of $u(n)$ from $q(n)$ because of the course quantisation: in the case of a 1-bit quantiser, $q(n)$ just represents the sign of $u(n)$. But $u(n)$ must have some relation with $x(n)$. From Fig. 1$a$ we obtain:

$$U(z) = \frac{G(z)}{1 + G(z)}[X(z) - E(z)]$$

which shows that $u(n)$ is highly correlated to $x(n)$. If estimates of $x(n)$ can be obtained, then they can be used to adapt the step size. It is easy to estimate the magnitude of $x(n)$ by lowpassing $q(n)$ and then finding the maximum magnitude over a certain period of time. According to the maximum-magnitude logic, it is reasonable to have the adaptive logic as follows

$$d_{i+1} = cM_i \quad d_{min} \leq d_{i+1} \leq d_{max} \tag{6}$$

where $M_i$ is the maximal magnitude in the $i$th block of the output samples of the lowpass filter, i.e. the maximum magnitude estimate of the input $x(n)$ for the $i$th block, and $d_{i+1}$ is the step size for the $(i + 1)$th block of samples. Each block contains, say, $K$ samples and $K$ depends on the stationary property of the signal. For speech, usually the signal can be considered stationary over 10–30 ms periods. For music signals, it may be less than 5 ms, i.e. $K$ should be less than 220, if the sampling frequency is 44.1 kHz. An appropriate initial value $M_0$ is required for both modulator and demodulator. In eqn. 6, $d_{max}$ depends on the maximum level of the system, $d_{min}$ depends on the requirement for noise level when the input is zero.

If the signal/noise ratio (SNR) for an A D system with no oversampling and noise-shaping is that SNR = $(\sigma_x^2/\sigma_e^2)$, where $\sigma_x^2$ is the signal power and $\sigma_e^2$ is the quantisation noise power, the total SNR for a sigma-delta modulator is

$$\text{SNR} = (\sigma_x^2/\sigma_e^2)\text{SNR}_{\text{enhancement}}$$

where $\text{SNR}_{\text{enhancement}}$ is obtained from oversampling and noise shaping techniques. For a sinusoidal input, $x = E \sin \omega_0 t$, $\sigma_x^2 = 0.5E^2$. In the case of no overload, and assuming that the quantisation noise is uniformly distributed in the range $|e| \leq d$, with the probability density function $p_e(e)$ being $1/2d$, then

$$\sigma_e^2 = \int_{-d}^{d} e^2 p_e(e)\,de = \int_{-d}^{d} \frac{e^2}{2d}\,de = \frac{d^2}{3}$$

Thus

$$\text{SNR} = (3E^2/2d^2)\text{SNR}_{\text{enhancement}}$$

Assuming the ideal case: $d_i = cE$ according to eqn. 6, then

$$\text{SNR} = (3/2c^2)\text{SNR}_{\text{enhancement}} \tag{7}$$

which means that SNR can be independent of the input level. From eqn. 7 it can be seen that the smaller the value of $c$ is, the better SNR can be obtained provided it does not cause overload.

The diagram of an adaptive sigma-delta modulator is shown in Fig. 4$a$ and the corresponding demodulator is

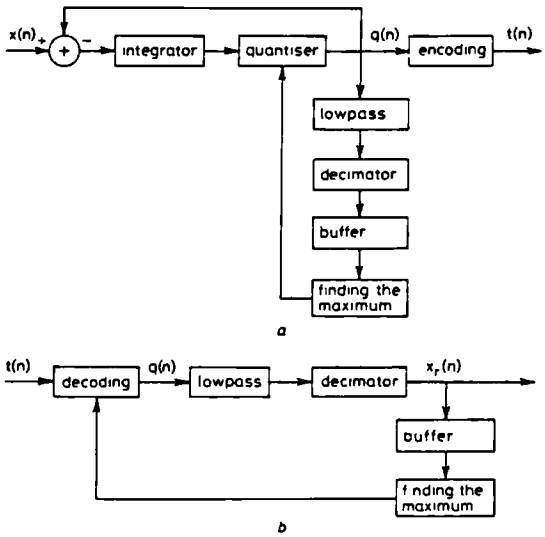



**Fig. 4** *Adaptive SDM*
*a* Modulator
*b* Demodulator

shown in Fig. 4$b$, where $t(n)$ is a digital sequence which is either 1 or 0 and $q(n)$ is an analogue sequence after decoding whose value changes according to the adaptation logic. Both modulator and demodulator use the same kind of lowpass filter and decimator. The main purpose of using lowpass filtering and decimation in the demodulator is to obtain a high quality for the reconstructed signal. The estimate of the magnitude of $x(n)$ can be obtained simultaneously. But in the modulator, the only purpose is to find the maximum magnitude of the signal so that a very sharp lowpass filter is not necessary. Thus a simple lowpass filter can be used in the modulator

but the complexity of the demodulator will be greater because two sets of lowpass and decimation systems have to be used.

One possible way of implementating the adaptation is by using a multiplying D A convertor (MDAC) in the feedback path as shown in Fig. 5. The output of MDAC

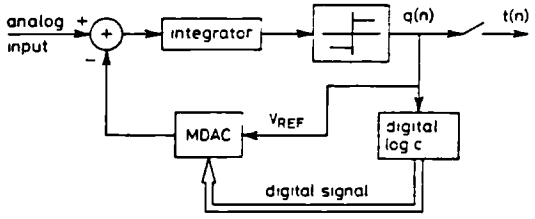



**Fig. 5** *Using MDAC to implement adaptation logic*

will be the result of multiplication of an analogue reference voltage (the output of the quantiser) and the output of the digital logic block.

For the lowpass filters and decimators in Fig. 4, the structure in Fig. 6 has been used. The comb filter in Fig.

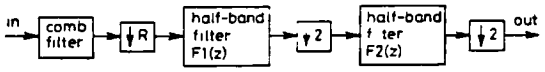



**Fig. 6** *Implementation of lowpass filter and decimator*

6 is a cascade of several simple comb filters $H_1(z)$

$$H_1(z) = \frac{1}{R}\left(\frac{1 - z^{-R}}{1 - z^{-1}}\right)$$

where $R = N/4$ and $N$, the oversampling ratio, is frequently a power of 2. The number of cascaded comb filters depends on the order of the sigma-delta modulator. For the $n$th-order SDM, $n + 1$ comb filters are cascaded [13]. The relation between the input and the output of the lowpass filter $H_1(z)$ is

$$y(n) = \frac{1}{R}\sum_{i=0}^{R-1} x(n - i)$$

which can be implemented easily in the feedback form

$$y(n) = y(n - 1) + (x(n) - x(n - R))/R$$

## 5 Simulation result

Adaptive quantisation has been carried out for 1-bit, 1st, 2nd, and 3rd-order sigma-delta modulators. For a 1-bit SDM, when the order is higher than 2, the system in eqn. 4 will become unstable in practice, mainly because the 1-bit loop quantiser is frequently overloaded, which is reflected by an increase in the amount of quantisation noise. This excess noise is circulated through the loop and can cause an even larger signal to appear at the quantiser input, eventually causing instability. A number of higher-order structures different from eqn. 4 have been presented which are stable [7, 14]. The structure shown in Fig. 7 has been used for simulations. The coefficients used are listed in Table 1.

**Table 1: Coefficients in the sigma-delta modulator**

| | $a$ | $b1$ | $b2$ | $b3$ |
|---|---|---|---|---|
| 1st-order* | 0 0 | 1.0 | 0 0 | 0.0 |
| 2nd-order* | 0.0 | 2.0 | 1.0 | 0 0 |
| 3rd-order† | 0.0011 | 1.0 | 0 5 | 0 1301 |

* from eqn. 4
† from computer simulations

For each stage in Fig. 7, the clipper is different because its input level becomes higher and higher as the stage number increases. The clipper of stage $j$ is defined as follows

$$y_j = \begin{cases} F_j & x_j > F_j \\ x_j & \text{elsewhere} \\ -F_j & x_j < -F_j \end{cases}$$

where

$$F_j = 3(1.2)^{j-1}d \quad (j = 1, 2, 3)$$

This result, which is considered optimal, is based on the computer simulations, where $d$ is the step size of the quantiser. The structure for the demodulator is the same as in Fig. 6, where the numbers of the comb filters cascaded are 2, 3, and 4, respectively, for the 1st-, 2nd-, and 3rd-order SDM, and 37th- and 129th-order FIR filters are used for the halfband filters $F1(z)$ and $F2(z)$ in Fig. 6, respectively.

In evaluating the sample spectrum of the reconstructed signal $x_r(n)$, the DFT $X_r(k)$ is calculated. A window function is applied to the sequence before the Fourier transform is taken. The window is as follows [15]

$$w(n) = 0.338946 + 0.481973\cos(\pi n/N) + 0.161054\cos(2\pi n/N) + 0.018027\cos(3\pi n/N)$$
$$n = -N, \ldots, -1, 0, 1, \ldots, N$$

which has a continuous third derivative and the first two equal side-lobes have a value of $-82.60$ dB. The length of the FFT is 1024.

The simulation for values of $c$ in eqn. 6 has been carried out when the value of $a$ is zero in Fig. 7, and results are shown in Fig. 8. It can be seen that $c$ must be greater than 1.0 for all 3 modulators, which means that

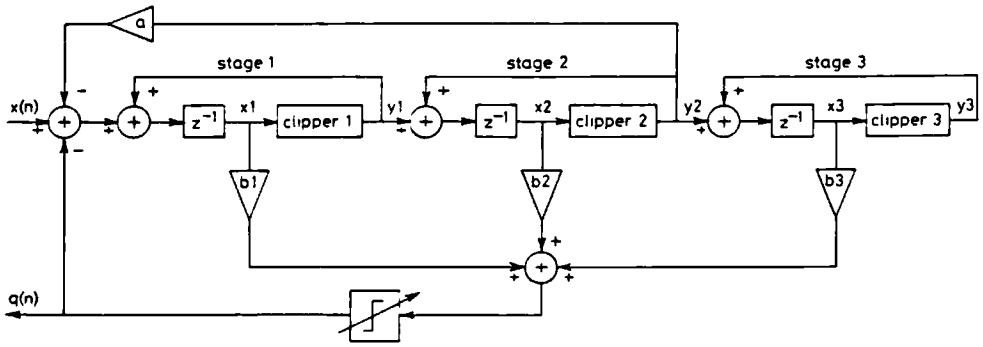



**Fig. 7** *Structure of a sigma-delta modulator: 1st-, 2nd-, or 3rd-order, depending on the coefficients*

the quantisation level should always be higher than the input level to avoid overloading the system. It also can be seen that for a certain range of values the systems are

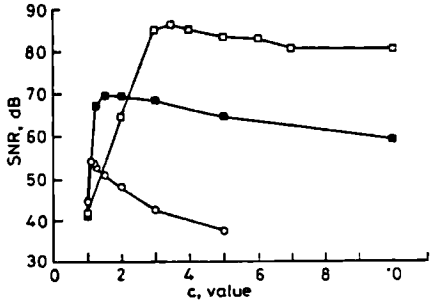



**Fig. 8** *Simulation results for varying c*
—○— 1st-order
—■— 2nd-order
—□— 3rd-order

not very sensitive to a change in $c$. Based on the results above, the following values are chosen

$$c = \begin{cases} 1.125 & \text{1st-order} \\ 1.5 & \text{2nd-order} \\ 3.5 & \text{3rd-order} \end{cases}$$

Considering that, for a music signal, the stationary time may be less than 5 ms, and the controlling factor $M_i$ for the current block of samples is calculated from the previous block, $K = 60$ is chosen as the block size, which corresponds to 1.36 ms. This $K$ value could be changed according to the statistical characteristics of the input, but is not investigated here. For fixed quantisers, assuming that $M$ is the maximum input level which does not cause overload, $cM$ is chosen as the quantisation level. The input $x(n)$ to the modulator and the output $x_r(n)$ of the demodulator are discrete time analogue signals, i.e. in computer simulation, floating point numbers are used to represent them.

The SNR results of both fixed and adaptive 3rd-order sigma-delta modulation, as the input level changes from 0 dB (maximum) to −60 dB, are shown in Fig. 9, which

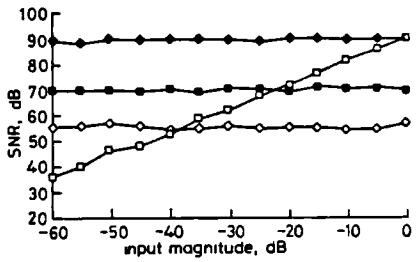



**Fig. 9** *SNR results for fixed and adaptive 3rd-order sigma-delta modulations, the 1st- and 2nd-order adaptive sigma-delta modulations*
input: 10 kHz sinewave
oversampling ratio: 64
Nyquist sampling frequency: 44.1 kHz
1-bit quantiser
—□— 3rd-order fixed SDM
—◆— 3rd-order adaptive SDM
—■— 2nd-order adaptive SDM
—◇— 1st-order adaptive SDM

clearly shows that the dynamic range of the system can be improved effectively by using an adaptive quantiser. Fig. 10 gives the spectrum results when the input level is −60 dB. Fig. 10*a* is the spectrum of the reconstructed signal when using a fixed quantiser and Fig. 10*b*, when an adaptive quantiser is applied. Fig. 9 also gives the results of the 1st- and 2nd-order adaptive sigma-delta modulations. They all have the same property: SNR is nearly independent of input level, which is consistent with the result from eqn 7. Fig. 11 gives the results of fixed and

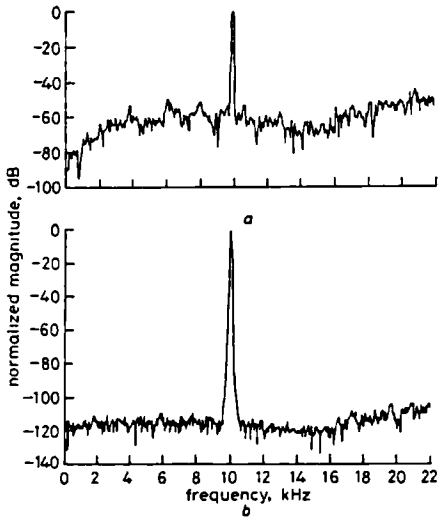



**Fig. 10** *Comparison of spectra of reconstructed signals between fixed and adaptive 3rd-order SDMs when input level is −60 dB*
*a* Using fixed quantiser
*b* Using adaptive quantiser

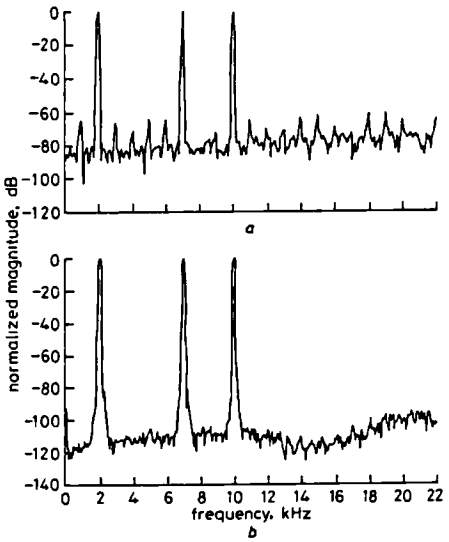



**Fig. 11** *Comparison of spectra of reconstructed signals between fixed and adaptive 3rd-order SDMs when input contains 3 tones at frequencies: 2, 7, and 10 kHz and the total input level is 10 dB*
*a* Using fixed quantiser
*b* Using adaptive quantiser

adaptive 3rd-order sigma-delta modulators when the input contains three tones whose total level is +10 dB which represents the overload case for a fixed SDM. Fig. 11*a* shows the spectrum of the reconstructed signal of the fixed SDM, from which it can be seen that, because of overload, the effects of harmonic and intermodulated components are very severe. However, in Fig. 11*b*, the harmonic and intermodulation distortion is reduced

effectively by allowing the quantisation level to increase as the input magnitude becomes larger. The difference between the signal level and the highest noise level in decibels can be seen from the spectra. However, this difference is not the SNR which is the result of the ratio of signal and noise integrals along the frequency or time axes. The SNR value is worse than this difference, which is obvious when comparing the spectra in Fig. 10 with the SNR results in Fig. 9.

It should be noted that eqn. 7 represents the ideal situation for a sinusoidal input. In practice, nonideal situations may occur, and they can be considered as a kind of interference on the value of $c$ in eqn. 7. Suppose that $c$ changes to $c + \Delta c$ and the derivative of SNR with respect to $c$ in eqn. 7 is

$$d\text{SNR}/dc = (-3/c^3)\text{SNR}_{\text{enhancement}}$$

so that

$$|\Delta\text{SNR}| \cong (3/c^2)|\Delta c\, c|\text{SNR}_{\text{enhancement}}$$
$$= 2|\Delta c/c|\text{SNR}_{\text{ideal}}$$

Considering the worst case

$$\text{SNR}_{\text{real}} = \text{SNR}_{\text{ideal}}(1 - 2|\Delta c/c|)$$

Therefore

$$\Delta\text{SNR(dB)} = 10 \log (1 - 2|\Delta c/c|)$$

In the case of $\Delta c/c$ being 0.01 and 0.1, $\Delta$SNR(dB) is $-0.088$ dB and $-0.97$ dB, respectively, which means that if $c$ changes by 10%, the loss in SNR will be about 1 dB.

## 6 Conclusions

It seems that adaptive sigma-delta modulation is a very promising technique: it gives a much wider dynamic range than a fixed quantiser. By defining the minimum step size $b_{min}$ appropriately, the idle channel noise can be reduced to a very low level. These advantages may give it many applications. For example, an adaptive sigma-delta modulator-demodulator may be directly connected to ADPCM coding without first converting it into linear PCM, thus the dynamic range of such a system will not be affected. When it is converted into linear PCM, although the dynamic range will be affected, some initial investigations show that it can still be used in some applications to reduce the oversampling ratio or the order of the loop filter while keeping the same quality for low level inputs. Considering the three main factors in a SDM: oversampling ratio, loop filter, and quantiser, the complexity among them always involves a trade-off. It depends on the applications as to which one is more important.

Future work will cover the following three areas. First more investigation on the parameter $K$ for different input signals and the effects of attack and decay in the adaptive quantiser, which may lead to a windowed version of $K$ which may smooth a sudden change caused by attack and decay. Secondly the effects on dynamic range of conversion to PCM, APCM, or ADPCM will be carried out. Thirdly, real-time implementation on a dedicated DSP chip will need to be studied. As mentioned before, a noise shaper plays the same role as sigma-delta modulator in reshaping the noise. Therefore, similar results and conclusions should be obtained from an adaptive noise shaper and this will also be investigated in the future.

## 7 References


1 INOSE, H., and YASUDA, Y.: 'A unit bit coding method by negative feedback', *Proc. IEEE*, 1963, pp. 1524–1535
2 CANDY, J.C.: 'A use of double integration in sigma-delta modulation', *IEEE Trans.*, 1985, COM-33, pp. 249–258
3 GRAY, R.M.: 'Oversampled sigma-delta modulation', *IEEE Trans.*, 1987, COM-35, pp. 481–488
4 UCHIMURA, K., HAYASHI, T., and IWATA, A.: 'Oversampling A-to-D and D-to-A converters with multistage noise shaping modulators', *IEEE Trans.*, 1988, ASSP-36, pp. 1899–1905
5 HAUSER, M.W.: 'Overview of oversampling A/D conversion', 89th AES Convention, Los Angeles, 21-25 Sept. 1990, Preprint 2973 (G-1)
6 WONG, P.W., and GRAY, R.M.: 'Two-stage sigma-delta modulation', *IEEE Trans.*, 1990, ASSP-38, pp. 1937–1952
7 CHAO, K.C.-H., NADEEM, S., LEE, W.L., and SODINI, C.G.: 'A higher order topology for interpolative modulators for oversampling A/D converters', *IEEE Trans.*, 1990, CAS-37, pp. 309–318
8 RABINER, L.R., and SCHAFER, R.W.: 'Digital processing of speech signals' (Prentice-Hall, Englewood Cliffs, N.J., 1978)
9 CARTIMALE, A.A.: 'Calculating the performance of syllabically companded delta-sigma modulators', *Proc. IEE*, 1970, 117, pp. 1915–1921
10 FLOOD, J.E., and HAWKSFORD, M.J.: 'Adaptive delta-sigma modulation using pulse grouping techniques'. Joint Conf. on Digital Processing of Signals in Communications, University of Technology, Loughborough, April 1972, pp. 445–462
11 TEWKSBURY, S.K., and HALLOCK, R.W.: 'Oversampled, linear predictive and noise-shaping coders of order $N > 1$', *IEEE Trans.*, 1978, CAS-25, pp. 436–447
12 JAYANT, N.S., and NOLL, P.: 'Digital coding of waveforms' (Prentice-Hall Inc., Englewood Cliffs, N.J., 1984)
13 CANDY, J.C.: 'Decimation for sigma delta modulation', *IEEE Trans.*, 1986, COM-34, pp. 72–76
14 RITONIEMI, T., KAREMA, T., and TENHUNEN, H.: 'Design of stable high order 1-bit sigma-delta modulators', 1990, IEEE Int. Symp. on Circuits and Systems, pp. 3267–3270
15 NUTTALL, A.H.: 'Some windows with very good sidelobe behavior', *IEEE Trans.*, 1981, ASSP-29, pp. 84–91

# BIBLIOGRAPHY

[1] K.W. Cattermole, *Principle of Pulse Code Modulation*, Iliffe Books Ltd., London, 1969.

[2] M.R. Aaron, "The Digital (R)Evolution", *IEEE Communications Magazine*, pp.21-22, January 1979.

[3] H. Inose, Y. Yasuda, and J. Murskani, "A Telemetering System by Code Modulation, Delta-Sigma Modulation", *IRE Trans. Space, Electronics, and Telemetery*, vol. **SET-8**, pp.204-209, Sept. 1962.

[4] H. Inose, and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback", *Proc. IEEE*, vol.**51**, pp.1524-1535, Nov. 1963.

[5] B.P. Agrawal, and K. Shenoi, "Design Methodology for EDM", *IEEE Trans. Commun.*, vol. **COM-31**, pp.360-370, March 1983.

[6] J.C. Candy, "A use of Limit Cycle Oscillations to Obtain Robust Analog-to-Digital Converters", *IEEE Trans. Commun.*, vol. **COM-22**, pp.298-305, March 1974.

[7] J.C. Candy, and O.J. Benjamin, "The Structure of Quantization Noise from Sigma-Delta Modulation", *IEEE Trans. Commun.*, vol. **COM-29**, pp.1316-1323, Sept 1981.

[8] J.C. Candy, "A Use of Double Integration in Sigma Delta Modulation", *IEEE Trans. Commun.*, vol. **COM-33**, pp.249-258, March 1985.

[9] J.C. Candy, "Decimation for Sigma Delta Modulation", *IEEE Trans. Commun.*, vol. **COM-34**, pp.72-76, Jan. 1986.

[10] R.M. Gray, "Oversampled Sigma-Delta Modulation", *IEEE Trans. Commun.*, vol. **COM-35**, pp.481-489, May 1987.

[11] R.M. Gray, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with DC input", *IEEE Trans. Commun.*, vol. **COM-37**, pp.588-599, June 1989.

[12] R.M. Gray, W. Chou, and P.W. Wong, "Quantization Noise in Single-Loop Sigma-Delta Modulation with Sinusoidal Inputs", *IEEE Trans. Commun.*, vol. **COM-37**, pp.956-967, Sept. 1989.

[13] Y. Matsuya, K. Uchimura, A. Iwata, T. Kobayashi, M. Ishikawa, and T. Yoshitome, "A 16-bit oversampling A-to-D Conversion Technology Using Triple-Integration Noise Shaping", *IEEE Journal of Solid-State Circuits*, vol. **22**, pp.921-929, Dec. 1987.

[14] K. Uchimura, T. Hayashi, T. Kimura, and A. Iwata, "Oversampling A-to-D and D-to-A Converters with Multistage Noise Shaping Modulators", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. **ASSP-36**, pp.1899-1905, Dec. 1988.

[15] K.C.-H. Chao, S. Madeem, W. Lee, and C.G. Sodini, "A Higher Order Topology for Interpolative Modulators for Oversampling A/D converters", *IEEE Trans. Circuits Sys.*, vol. **CAS-37**, pp. 309-318, Mar. 1990.

[16] B.P. Brandt, D.E. Wingard, and B.A. Wooley, "Second-Order Sigma-Delta Modulation for Digital-Audio Signal Acquisition", *IEEE Journal of Solid-State Circuits*, vol. **26**, pp.618-627, April 1991.

[17] R.W. Adams, P.F. Ferguson Jr., S. Vincellette, A.Ganesan, T.Volpe, and B. Libert, "Theory and Practical Implementation of a 5th-order Sigma-Delta A/D Converter", *the 90th Convention of Audio Engineering Society* , Preprint **3017 (C-2)**, Paris, Feb. 1991.

[18] P.W. Wong, and R.M. Gray, "FIR Filters with Sigma-Delta Modulation Encoding", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. **ASSP-38**, pp.979-990, June 1990.

[19] R. Nawrocki, and J. Yu, "Waiting Time Jitter Reduction by Sigma Delta Modulation", to be submitted to IEE proceedings-I.

[20] D.M. Green, *An Introduction to Hearing*, Hillsdale, New Jersey: Lawrence Eelbaum Assoc., Inc., 1976.

[21] T. Muraoka, M. Iwahara, and Y. Yamada, "Examination of audio-bandwidth requirements for optimum sound signal transmission", *J. Aud. Eng. Soc.*, vol. **29**, no. 1/2, pp. 2-9, Jan./Feb. 1982.

[22] T. Muraoka, Y. Yamada, and M. Yamazaki, "Sampling-frequency considerations in digital audio", *J. Aud. Eng. Soc.*, vol. **26**, no. 4, pp. 252-256, April 1978.

[23] L.R. Fincham, "The subjective importance of uniform group delay at low frequencies", *The 74th Convention of Audio Engineering Society* , Preprint **2056 (H-1)**, New York, Oct. 1983.

[24] L.D. Fielder, "Dynamic-range requirement for subjectively noise-free reproduction of music", *J. Aud. Eng. Soc.*, vol. **30**, no. 7/8, pp. 504-511, July/Aug. 1982.

[25] R. Steele, *Delta Modulation Systems*, Pentech Press Ltd., 1975.

[26] M.S. Ghausi and K.R. Laker, *Modern Filter Design*, Prentice-Hall, 1981.

[27] A. Gersho, "Principles of Quantization", *IEEE Trans. Circuits and Systems*, vol. **CAS-25**, pp.427-436, Jul. 1978.

[28] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984.

[29] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[30] S.K. Tewksbury, and R.W. Hallock, "Oversampled, Linear Predictive and Noise-Shaping Coders of Order N>1", *IEEE Trans. Circuits and Systems*, vol. **CAS-25**, pp. 436-447, Jul. 1978.

[31] E.F. Stikvoort, "Some Remarks on the Stability and Performance of the Noise Shaper or Sigma-Delta Modulator", *IEEE Trans. Commun.*, vol. **COM-36**, pp.1157-1162, Oct. 1988.

[32] M.W. Hauser, "Overview of Oversampling A/D conversion", *the 89th AES Convention*, Preprint **2973 (G-1)**, Los Angeles, Sept. 1990.

[33] S. Hein and A. Zakhor, "Stability and Scaling of Double Loop ΣΔ Modulators", in *Proc. IEEE Int. Symp. Circuits and Systems*, pp.1312-1315, San Diego, USA, May 1992.

[34] V. Friedman, "The Structure of the Limit Cycles in Sigma Delta Modulation", *IEEE Trans. Commun.*, vol. **COM-36**, pp.972-979, Aug. 1988.

[35] E. Dijkstra, L. Cardoletti, O. Nys, C. Piguet, and M. Degrauwe, "Wave Digital Decimation Filters in Oversampled A/D Converters", in *Proc. IEEE Int. Symp. Circuit and Systems*, vol.**3**, pp.2327-2330, 1988.

[36] T. Saramaki, T. Karema, T. Ritoniemi, and H. Tenhunen, "Multiplier-Free Decimator Algorithms for Superresolution Oversampled Converters", in *Proc. IEEE Int. Symp. Circuit and Systems*, vol.**4**, pp.3275-3278, New Orleans, LA, USA, May 1990.

[37] T. Hayashi, Y. Inabe, K. Uchimura, T. Kimura, "A multistage delta-sigma modulator without double integration loop", *ISSCC Digest of Technical Papers*, pp.182-183, Feb. 1986.

[38] T. Ritonniemi, T. Karema, and H. Tenhunen, "Design of Stable High Order 1-Bit Sigma-Delta Modulators", in *Proc. IEEE Int. Symp. Circuit and Systems*, vol.4, pp.3267-3270, New Orleans, LA, USA, May 1990.

[39] CCITT Recommendation G.712, "Performance Characteristics of PCM Channels Between 4-Wire Interfaces at Voice Frequencies", Blue Book, Vol.III, 1989.

[40] R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[41] J.-J. E. Slotine, W. Li, *Applied Nonlinear Control*, Prentice-Hall, Inc., 1991.

[42] E. Kreyszig, *Advanced Engineering Mathematics*, John Wiley & Sons, Inc., 1983.

[43] S. Hein and A. Zakhor, "On the Stability of Interpolative Sigma Delta Modulators", in *Proc. IEEE Int. Symp. Circuits and Systems*, pp.1621-1624, Singapore, Jun., 1991.

[44] O. Feely, and L.O. Chua, "The effect of Integrator Leak in Sigma-Delta Modulation", *IEEE Trans. Circuits and Syst.*, vol. **CAS-38**, pp. 1293-1305, Nov. 1991.

[45] A. A. Cartmale, "Calculating the performance of syllabically companded delta-sigma modulators", *Proc. IEE*, 1970, vol. **117**, pp. 1915-1921, Oct. 1970.

[46] J. E. Flood, and M. J. Hawksford, "Adaptive delta-sigma modulation using pulse grouping techniques", *Joint Conference on Digital Processing of Signals in Communications*, University of Technology, Loughborough, pp. 445-462, April, 1972.

[47] A.H. Nuttall, "Some Windows with Very Good Sidelobe Behavior", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. **ASSP-29**, pp.84-91, Feb. 1981.

[48] R.J. Tocci, *Digital Systems, Principles and Applications*, Fourth Edition, Prentice-Hall, Inc. 1988.

[49] C.R. Caine, A.R. English, and J.B. O'Cleary, "Nicam 3: near-instantaneously companded digital transmission system for high quality sound programmes", *The Radio and Electronic Engineer*, vol. **50**, pp. 519-530, Oct. 1980.

[50] E. A. Lee, and D. G. Messerschmitt, *Digital Communication*, Chapter 17, Kluwer Academic Publishers, Boston, 1988.

[51] A.V. Oppenheim, and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall International Inc., 1975.

[52] B.E. Boser, K.-P. Kartmann, H. Martin, and B.A. Wooley, "Simulating and Testing Oversampled Analog-to-Digital Converters", *IEEE Trans. Computer-Aided Des.*, vol. **CAD-7**, pp.668-674, June, 1988.

[53] CCITT Recommendation O.132, "Quantizing Distortion Measuring Equipment Using a Sinusoidal Test Signal", Blue Book, IV.4, 1989.

[54] P.R. Adby, and M.A.H. Dempster, *Introduction to Optimization Methods*, Chapman and Hall Ltd., 1978.

[55] P.J. Bloom, "High-Quality Digital Audio in the Entertainment Industry: An Overview of Achievements and Challenges", *IEEE ASSP Magazine*, pp.2-25, Oct. 1985.

[56] J.M. Goldberg, and M.B. Sandler, "Noise Shaping and Pulse-Width Modulation for an All-Digital Audio Power Amplifier", *J. Audio Eng. Soc.*, vol. 39, pp.449-460, June, 1991.

[57] A. Paul, "*Simulate* (version 3.0): User's Manual", DSP Lab, Department of Electronic and Electrical Engineering, King's College London, Feb. 1992.

[58] *Programs for Digital Signal Processing*, Edited by Digital Signal Processing Committee, IEEE ASSP Society, IEEE Inc., 1979.